IN THE UNITED STATES DISTRICT COURT

FOR THE NORTHERN DISTRICT OF ILLINOIS

EASTERN DIVISION

STACY ERNST, DAWN HOARD, KATHERINE KEAN, MICHELLE LAHALIH, and IRENE RES-PULLANO Plaintiffs, vs. CITY OF CHICAGO, Defendant.	Case No.: 08CV4370 Judge Charles R. Norgle, Sr. Magistrate Judge Jeffrey Cole

PLAINTIFFS' PROPOSED FINDINGS OF FACT AND CONCLUSIONS OF LAW

TABLE OF CONTENTS

Page

I. PROPOSED FINDINGS OF FACT[1]

1. This case has been brought by five women—Stacy Ernst, Dawn Hoard, Irene Res-Pullano, Michelle Lahalih, and Katherine Kean—who applied for positions as paramedics with the Chicago Fire Department ("CFD") and were denied those positions due to their scores on a pre-employment Physical Performance Test ("the PPT") administered by the City of Chicago ("the City") to eligible applicants. The Plaintiffs filed timely charges with the EEOC and have brought their suit under Title VII, advancing two theories of liability: (1) that the City intentionally discriminated against women by its use of this test, and (2) that the City's use of the test had a prohibited disparate impact against women.[2]

2. In November 2014, the parties tried Plaintiffs' claim of intentional discrimination (or "disparate treatment") to a jury, which returned a verdict for the City. Although the evidence bearing on the intent and disparate impact claims overlaps, because the elements of proof for the two claims are significantly different and because a plaintiff may demonstrate a violation of Title VII under the disparate impact theory without proving discriminatory intent, the jury's finding of no intentional discrimination does not dictate or suggest the result of Plaintiffs' still pending Title VII disparate impact claim. The jury, for example, could have believed the PPT was manifestly invalid, yet found for the City by deciding that use or continued use of the test was not motivated by discriminatory intent. *Melendez v. Ill. Bell Tel. Co.*, 79 F.3d 661, 670 (7th Cir. 1996).

[1] Any Finding of Fact appearing below that should more fittingly be denominated as a Conclusion of Law should be given that effect, and vice versa.

[2] It is well established that the disparate impact theory can be used not only as the basis for a class action but also form the basis of an individual claim. *Stockwell v. City and County of San Francisco*, 749 F.3d 1107, 1115 (9th Cir. 2014) (and cases cited therein); *see also*, *Melendez v. Ill. Bell Tel. Co.*, 79 F.3d 661 (7th Cir. 1996) (addressing an individual disparate impact claim)

II. BACKGROUND

3. For many years prior to April 2000, applicants applied for positions as paramedics with the Chicago Fire Department by submitting a completed application and documentation verifying a valid Illinois Driver's License, a current Illinois EMT/Paramedic license, residency in the City of Chicago, high school graduation, selective service registration (if applicable), and a birth certificate. *See* PX 14A at ERN002861.[3] The City would review these materials to determine their sufficiency, and the names of applicants whose documentation was in order were placed on an "eligibility list." Relative position on the list was determined on a seniority basis, determined by the date of application. *Leen v. Carr*, 945 F. Supp. 1151, 1153-54 (N.D. Ill 1996) (describing the City's paramedic application and hiring process). As paramedic positions opened, the City filled them by selecting names off the eligibility list, proceeding from the top down. Those selected were made conditional offers of employment, which ripened into full offers upon satisfactory completion of a drug screen, medical exam, and background investigation. With minor variations, this hiring system appears to have been in place from the mid 1970's through the 1990's. During these years, no applicant for a paramedic position with the CFD was ever required to take or pass a pre-employment physical fitness or performance test as a condition of employment. *See* Tr. at 970:6-24, 971:20, 1183:23-1184:15; 1187:10-20, 1432:16-18, 1441:6-8, 1442:22-1443:11; PX 14A at ERN002861; *see also Leen*, 945 F. Supp. at 1153-54.

[3] Throughout these findings, the following abbreviations are used. "PX" refers to plaintiffs' Exhibits, "DX" to defense exhibits, and Bates numbers following a PX or DX designation (for example, "ERN000__") to the precise Bates page of any multi-page exhibit where the cited evidence can be found. "Tr." refers to the trial transcript. "CFD" refers to the Chicago Fire Department. "PPT" refers to physical performance test, administered to applicants for paramedic positions with the CFD, that is the subject of this case.

4. Throughout the 1980's and 1990's, the City hired paramedics through this process. As of October of 1999, the CFD was employing 296 paramedics, 172 paramedics in charge, and 59 ambulance commanders, the vast majority of whom had been hired by the CFD through this process. DX 37 at ERN000582; PX 14C at ERN0002769. They were hired without ever submitting to or passing a pre-employment physical fitness or performance test.

5. The City introduced no competent evidence that this time-tested method of hiring did not meet the CFD's legitimate needs. The record, for example, contains no competent evidence—documentary or testimonial, objective or anecdotal—that any patient was ever jeopardized by any paramedic hired by the City without physical performance testing—or even that any patient's transport was ever delayed. The record contains no evidence that any CFD paramedic hired without taking and passing a physical performance test ever performed unsatisfactorily in any respect due to a lack of physical fitness or ability. The record contains no evidence that any CFD paramedic has ever been suspended, disciplined, demoted, or terminated due to either an inability to perform, or difficulty in performing, any of the physical requirements of the paramedic job. The record contains no evidence that women's job performance differed from men's before the introduction of the PPT in April 2000. The record also contains no evidence that paramedic job performance has improved on any measure of performance since the implementation of the PPT in April 2000. Tr. at 1454:1-5. After years of discovery and litigation and a multi-week trial, these are glaring holes in the evidence—on an issue (job-relatedness) on which the City bore the burden of proof.

6. The issue in this case is not and has never been whether the CFD must jeopardize patient safety in order eliminate disparate impact against female applicants. Rather, the point is that patient safety is not at issue. There is no evidence of any risk to patient safety. The City,

which bears the burden of proof on the issue of job relatedness, has introduced no competent evidence to establish a safety-based job-relatedness or business necessity defense. It also bears noting that any such defense would have required the City to introduce hard evidence: job-relatedness and business necessity "cannot be proved through vague and unsubstantiated hearsay." *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 428 n.23 (1975).

7. In April 2000, the City radically changed its process for hiring paramedics, by instituting a compulsory physical performance test for all applicants for paramedic positions. Overnight, this new requirement erected a barrier to employment for 40% of the women applying for the job, but only 2% of men. Women were, on average, twenty times more likely than men to be eliminated in the hiring process by the PPT. Tr. at 625:1-10, 626:5-631:16, and 1654:24-1656:13; PX 21 at pp. 14-19; PX 22 at pp. 12-13.

8. Between 2000 and 2009, the City administered the PPT six times, to some 1,088 applicants for CFD paramedic positions. Over this period, the passing rate for women was 59.18%. By contrast, the passing rate for men was far higher, 98.24%. Tr. at 627:19-629:2; PX 21 at p. 17. The passing rate for women was, accordingly, only 60% of the passing rate for men, well below the 80-percent standard set by the EEOC to implement the disparate-impact provision of Title VII, as part of its Uniform Guidelines on Employee Selection Procedures. 29 C.F.R. § 1607.4D; *Ricci v. DeStefano*, 557 U.S. 557, 586-87 (2009); PX 21 at p. 17.[4]

[4] Guidance with respect to standards and principles for the development and validation of employment tests has been offered by various governmental and professional bodies in four documents: (a) the Uniform Guidelines on Employee Selection Procedures ("the Uniform Guidelines") (1978), 29 C.F.R. pt. 1607; (b) the accompanying Questions and Answers on the Uniform Guidelines on Employee Selection Procedures ("EEOC Questions & Answers"), 44 Fed. Reg. 11,996 (Mar. 2, 1979), jointly promulgated by the EEOC, Civil Service Commission, Department of Labor and Department of Justice; (c) the Principles for the Validation and use of Personnel Selection Procedures ("the Principles") (1987), promulgated by Division 14 of the American Psychological Association, the Society for Industrial and Organizational Psychology ("SIOP"); and (d) the Standards for Educational and Psychological Testing ("the Standards"),

9. This steep disparity was statistically significant: the pass rates of men and women were separated from the results to be expected from a sex-neutral selection rate by more than 17 units of standard deviation. PX 21 at pp. 17-20. In August 2004, when the five plaintiffs in this case took the physical performance test, the disparate impact was even greater: the passing rate for women was just 49% of the passing rate for men; for that administration, the pass rates of men and women were separated from the results to be expected from a sex-neutral selection rate by more than nine standard deviations. PX 21 at p. 18.[5]

(continued …)

jointly promulgated by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education.

[5] The results and adverse-impact ratios for the six administrations of the PPT between 2000 and 2009 are reproduced below (from PX 21 at p. 18):

Table 1: Adverse Impact by Gender for Physical Abilities Tests for Chicago Paramedics

Passing Rates by Administration and Gender									
	Women			Men			Grand Totals		
Admin	Number Passing	Number of Candidates	Pass Rate	Number Passing	Number of Candidates	Pass Rate	Number Passing	Number of Candidates	Pass Rate
All	174	294	59.18%	780	794	98.24%	954	1,088	87.68%
April 2000	42	67	62.69%	159	160	99.38%	201	227	88.55%
June 2001	46	78	58.97%	176	180	97.78%	222	258	86.05%
August 2004	34	73	46.58%	172	180	95.56%	206	253	81.42%
July 2005	29	46	63.04%	156	157	99.36%	185	203	91.13%
October 2008	13	18	72.22%	74	74	100.00%	87	92	94.57%
March 2009	10	12	83.33%	43	43	100.00%	53	55	96.36%

Adverse Impact Ratios and Standard Deviation Differences by Administration and Gender		
	Women	
Admin	Adverse Impact Ratio	Standard Deviation Difference
All	0.60	-17.41
April 2000	0.63	-7.92
June 2001	0.60	-8.26
August 2004	0.49	-9.08
July 2005	0.63	-7.62
October 2008	0.72	-4.66
March 2009	0.83	-2.73

10. Significantly, no CFD incumbent paramedic on the job was ever required to take a physical performance test between 2000 and 2009 and, as far as the record discloses, none ever has been. *See* Tr. at 528:16-19. The City does not require its incumbents to satisfy the standard it imposes on new hires. See Tr. at 1840:21-23. This, alone, undermines the City's claim that a physical performance test is job related or consistent with business necessity. If whatever level of physical performance is measured by the PPT were related to patient safety, or job-related and consistent with business necessity for purposes of Title VII, then it would be as necessary to continually measure physical performance later in a paramedic's career as it would be at the outset; indeed, perhaps even more so given the decrements of age.[6]

11. If the City had a gender-neutral rationale for instituting the PPT in April 2000, it remains shrouded in mystery.

12. The City has introduced *no* evidence of *any* observed difference between men and women in actual paramedic job performance. A veteran paramedic, promoted to the rank of ambulance commander who has conducted field training for new hires, testified that he has observed no difference in the physical abilities of candidates hired before and after the introduction of the PPT. Tr. at 1190:18-1191:1.

13. At trial, only one City witness, Charles Stewart, even attempted to offer any rationale for the PPT. He did not have any actual role in the development of the physical performance test. Tr. at 535:12. But he hypothesized that the rationale could have been a concern to reduce injuries among new paramedic hires during initial training. Tr. at 537:17-539:4. This

[6] The City's argument about the job relatedness of the PPT is also undermined by the fact that the PPT is administered to some applicants two years or more before they enter the training academy. Tr. 1243:1-6. Yet they are not required to re-test or demonstrate their continued fitness or ability before starting the academy. *Id.*

testimony was vague and unsubstantiated hearsay—"I was notified either by Training or the Medical section . . . [and] I reported [it] to my immediate supervisor." Tr. at 537:23-539:4. As such, this testimony is incompetent to prove job relatedness. *Albemarle Paper Co. v. Moody*, 422 U.S. at 428 n.23. It is an impermissible, inadequate stand-in for concrete evidence based on a job analysis.

14. The City has introduced no evidence of injury rates in the Academy (or on the job), before or after implementation of the PPT, and no evidence how the PPT was designed to, expected to, or whether it in fact had any effect on injury rates. Records of paramedics and paramedic in training who are injured and laid up are, as one would expect, created and maintained. Tr. at 1449:25-1450:2. One can only surmise that the City did not introduce them because they would have undermined Chief Stewart's hearsay testimony. DX 79 identifies paramedics who have separated from employment during Academy training since January 1990, for an entire decade before the introduction of the PPT and identifies no paramedic having separated due to injury. There is no competent evidence of injuries occurring in training laying up any paramedics. *See also* Tr. 973:11-974:8 (complete class graduated). The City introduced no data on paramedics' medical leave or absences either before or after the adoption of the PPT.[7]

15. There is no evidence in this record that any large city other than Chicago requires applicants for paramedic positions to take and pass a pre-employment physical performance test.

16. The idea to introduce a pre-employment physical abilities test for applicants for paramedic positions appears to have originated with outside consultants rather than stemming

[7] *More* paramedics failed to complete Academy training *after* the adoption of the PPT than before: the adoption of the PPT has decreased rather than increased success in training. DX 79. This occurred despite the fact that the CFD's goal was for the PPT was to have candidates who would succeed in training. DX 39 at p. 8.

from any demonstrated (or even perceived) need from within the CFD. The first indication comes from August 1995, when, having just completed work for the CFD on the physical assessment portion of a hiring exam for entry-level firefighters, Human Performance Systems (HPS), headed by Dr. Deborah Gebhardt, began urging the City to hire the firm for more work, proposing to develop and sell the City a PPT for the paramedic position. *See* DX 10; DX 11.

17. Significantly, Dr. Gebhardt's sales pitch to the City did not focus on risks to patient safety from hiring, or continuing to hire, paramedics without requiring them to take and pass a pre-employment physical performance test. Rather, she primarily advertised that *employee* safety might be improved by implementing pre-employment physical performance assessment. DX 11 at ERN004377. There is no competent evidence in the record in this case that either Dr. Gebhardt or the City ever proposed introducing, or in fact introduced, a pre-employment PPT due to any actual, or perceived, risk to patient safety.

18. By October 1996, the proposal HPS had submitted the year before had not been acted on. In October 1996, HPS submitted a revised proposal to the City, again urging the City to hire the firm to develop a pre-employment physical performance test. *See* DX 11. This time, HPS's marketing was successful.

19. In March 1997, Adrienne Bryant, Deputy Chief of Staff to the Mayor, requested that the City's Purchasing Department authorize a "sole source," or "no-bid," contract for HPS to develop and validate a pre-employment PPT for CFD paramedics, citing Dr. Gebhardt's "good working relationship with CFD's top Brass," which "serves as a strong incentive" for awarding the contract without putting the work out for bid. DX 12; Tr. at 562:11-14.

20. In January 1998, the City accepted HPS's proposal and hired the firm, without issuing an RFP or requesting or receiving competing proposals or bids. *See* DX 13; DX 14; DX

15. With a contract in hand, HPS began work developing and attempting to validate a physical performance test for the CFD paramedic position, starting with a "job analysis." *See* DX 16; DX 17.

III. THE JOB ANALYSIS

21. The purpose of a job analysis is to identify the essential tasks required to successfully perform the job being studied, along with the knowledges, skills and abilities required to carry out those tasks. Tr. at 616:6-10; DX 14 at ERN004224-27; 29 C.F.R. § 1607.16K.

22. Dr. Gebhardt and HPS conducted their job analysis of the CFD paramedic position in 5 stages, comprised of: (1) collecting job information through interviews of incumbents and direct observations of job performance; (2) generating an "inventory" of the "tasks" CFD paramedics perform; (3) asking incumbents to fill out a job-analysis questionnaire "rating" the relative importance of tasks on the task inventory; (4) analyzing the incumbents' ratings to isolate "essential" tasks; and, (5) asking incumbents to "link" the "essential" tasks to the physical abilities required to successfully carry them out. DX 37 at ERN000577-596.

A. Observation of the Job: Interviews and Ride-Alongs.

23. As a first step toward developing the PPT, Dr. Gebhardt reviewed information about the position and the existing hiring process, provided to them by the CFD, conducted structured interviews with CFD paramedic personnel, and, one evening, accompanied CFD paramedics on a "ride-along" to observe what CFD paramedics do, how they do it, the context in which they do it, and to try to gain an understanding of the knowledge, skills, abilities required to perform the job successfully. DX 37 at ERN000577-578.

B. The "Task" "Inventory."

24. In the second stage of the job analysis process, Dr. Gebhardt took the information generated by their interviews and observations, together with information from their own past studies of the paramedic position for other clients, and generated "task list," or "task inventory," of tasks performed by CFD paramedics. *See* DX 37(C).

25. This task inventory described the CFD paramedic job in terms of 156 tasks but was heavily weighted toward the physical tasks associated with the job. Of the 156 tasks listed, 66 were devoted to physical tasks. A full 22 were devoted to lifting alone. DX 37(C); PX 21 at pp. 22-23. By comparison, only 27 tasks were devoted to "technical" skills (like starting an IV, intubating, or performing CPR). DX 37 at ERN000685-686. The result gives a distorted view of a paramedic job, which overemphasizes and overvalues its physical aspects. PX 21 at pp. 22-23; Tr. at 641:3-15; 642:21-643:6. The job analysis was conducted in a manner that predisposed the result. Tr. 632:19-633:5.

26. One way to gauge the extent of Dr. Gebhardt's over-emphasis on the physical aspects of the paramedic job is by comparing her task inventory for the paramedic position to the one contained in the United States Department of Labor's Occupational Information Network ("O*NET"). *See* DX 117; Tr. at 645:20-646:23. The O*NET database is an independent and neutral source of comprehensive job descriptive information, developed by professional job analysts for the Department of Labor, based on information from job incumbents, occupational experts, and occupational experts. The O*NET is the Department of Labor's official source of job information. It identifies the knowledge, skills, abilities, and other occupational characteristics necessary for successful performance of more than 900 occupations. *See* U.S. Dep't of Labor, Employment & Training Admin., O*NET - beyond information – intelligence, *available at* www.doleta.gov/programs/onet/eta_default.cfm (last visited January 13, 2015); The

O*NET Data Collection Program, *available at* www.doleta.gov/programs/onet/datacollection.cfm (last visited January 13, 2015); O*NET FAQ, *available at* www.onetcenter.org/questions/10.html?c=Top (last visited January 13, 2015). Tr. at 645:20-646:23, 732:22-733:4; PX 22 at pp. 19-20.

27. For the paramedic position, the Department of Labor's Occupational Information Network identifies fifteen "core" or essential tasks. Only one of those fifteen requires significant physical exertion: "Immobilize patient for placement on stretcher and ambulance report, using backboard or other spinal immobilization device." DX 117 at p. 2. By contrast, physical tasks comprise 42% of the tasks on Dr. Gebhardt's task list. DX 37 at ERN000682-685; PX 21 at p.22.

28. Contributing to Dr. Gebhardt's over-emphasis on the physical aspects of the paramedic position, there are also a series of conspicuous omissions in her job analysis. The job analysis neither acknowledges nor takes into account that:

a. It is CFD policy to dispatch a fire engine, along with an ambulance, on most calls, including on all critical traumas, on all non-critical medical calls when the ambulance is more than 18 blocks away (PX 13 at ERN010955, PX 14A at ERN002838, PX 16) to the scene of all vehicular accidents (Tr. at 199:12-201:7), and in response to any call by paramedics for assistance with lifting (Tr. at 198:22-199:11);

b. Paramedics have assistance from other paramedics, EMTs, or firefighters on almost 80 percent of all emergency calls (DX 39 at p. 9); and,

c. Twenty-plus percent of all calls involve no transports. PX 14A at ERN002820.[8] Acknowledging the relevant of this kind of information, Dr. Gebhardt had

[8] ERN2820 shows 232,716 runs for calendar year 1997, of which 178,903 involved transport, leaving 23% where no transport was required. *See also* PX 50A-C (reports of unassisted calls without transport and pediatric calls).

represented that her job analysis would address "the presence/absence of assistance from coworkers (e.g. teamwork)." DX 11 at ERN004383. In fact, none of this information is addressed in Dr. Gebhardt's validation study. DX 37. Her job analysis ignores the fact that paramedics virtually always perform the physical tasks of the job with a partner and, often, with assistance from firefighters as well.

29. Dr. Gebhardt's job analysis also displays no awareness that paramedics spend their first months on the job in a training academy, where they participate in physical training on a daily or nearly daily basis, including training on using stair chairs, long boards and stretchers to lift and carry patients (Tr. at 1443:5, 1836:3-7); that at the end of their academy training they must demonstrate their ability to handle the physical demands of the job (Tr. at 543:6-24) including by carrying a dummy upstairs on a stretcher for two or three flights (Tr. at 1184:23-1185:14l; *see also* Tr. at 973:11-974:3, 1195:11-1196:2); or, that the level of physical ability measured by the PPT is not, therefore, indicative of the level of physical ability that paramedics need at the time of hire, as opposed to when they step into the field, given the undisputed evidence that physical performance is highly trainable. PX 21 at pp. 8, 24-25; Tr. at 453:13-16. This omission is particularly significant given that Dr. Gebhardt evaluated whether other job requirements (such as personality traits) were needed prior to training or only after training. DX 37D at ERN000715-718; *see also* PX 21 at pp. 24-25.

C. <u>The Job Analysis Questionnaire.</u>

30. In the third stage of their job analysis, Dr. Gebhardt asked a 100-plus CFD incumbent paramedics, paramedics in charge, and ambulance commanders to complete a job analysis questionnaire, which required them to try to assign numerical "ratings" to each of the

156 tasks listed in the task inventory, attempting to quantify "frequency," "importance," "time spent," and "physical effort." DX 37 at ERN000691.

31. After collecting the incumbents' ratings, Dr. Gebhardt used them to narrow down the list of 156 tasks. She narrowed the list to 24 essential "physical" tasks. *See* DX 37 at ERN000588. By thus focusing only on the "physical" tasks, Dr. Gebhardt considered only a small fraction of the job domain, excluding the vast majority of the job, and many of the most critical and important tasks of the job. Of the tasks she eliminated, many had been rated "very important" or "of great importance" by incumbents. On the task inventory of 156 tasks, incumbents had rated 78 of them as "very important or "of great importance." Dr. Gebhardt eliminated 64 of them from further consideration. *Compare*, e.g., DX 37C and DX 37K at ERN000756-757.

32. As part of this same exercise, incumbents were also given: (a) a set of "supplemental questions" that addressed "the weights of [patients and] equipment carried, distances equipment was carried, flights of stairs climbed, environmental working conditions and other job parameters;" and, (b) a list of personality characteristics, which they were asked to rate for their contribution to successful performance as a paramedic. DX 37 at ERN000579.

33. Dr. Gebhardt used the incumbents' answers to the supplemental questions regarding weights and distances as the sole basis for estimating of how often CFD paramedics lift heavy patients and how heavy those patients are. However, the underlying data is unreliable and the method by which it was collected is inconsistent with *Daubert* principles. Regarding the weight of patients, Dr. Gebhardt posed the following question to incumbents:

> Of the patients you lifted or lifted and carried in the past year, what <u>percent</u> of the patients weighed:
>
> Less than 90 lbs. _____ %
> 90 to 150 lbs. _____ %

160 to 200 lbs. _____ %
210 to 250 lbs. _____ %
60 to 300 lbs. _____ %
More than 300 lbs. _____ %

This methodology required incumbents to guesstimate the weights of patients. It was a self-evidently questionable means of gathering information, resulting in job analysis information of dubious, if any, accuracy. Tr. at 752:20-753:1, 737:15-18. Equally or more importantly, asking incumbents to guesstimate was demonstrably unnecessary. Better and more reliable *actual empirical* data existed or could have been gathered. The run sheets that CFD paramedics complete for each patient contain a space for patient weight. PX 14C at ERN002773. And, Dr. Gebhardt knew that, both from having received copies of the run sheet form from the CFD (*id.*) and from their own observations on their ride-alongs with CFD paramedics. Indeed, Dr. Gebhardt and HPS's own notes of the patients they encountered during that ride-along record the weight of each patient. *See* DX 37A at ERN000667-677. Obtaining actual patient weights was clearly feasible.

34. In view of the existence or availability of actual empirical data regarding patient weights, Dr. Gebhardt's refusal to use or obtain it, in favor of asking incumbents to speculate about patient weights instead, was slapdash, unscientific, and unreliable. And, the results this approach yielded were predictably erratic. One respondent reported that patients weigh 260-300 pounds 45% of the time and over 300 pounds 50% of the time. By contrast, more than a third of the respondents (35 out of 110) reported 0 patients over 300 pounds. PX 22 at p. 14 n.2. *See also* Tr. at 737:13-20.

35. The responses Dr. Gebhardt collected to the supplemental questions also suffer from a second fatal flaw. All data collection methods "result in information with some error." DX 37 at p. 8. For that reason, professional and legal standards call for providing estimates of the

level of error by reporting reliability coefficients. *Gillespie*, 771 F.2d at 1041 (a criterion-related test "must be tested for reliability"); *United States v. State of Delaware*, No. 01-cv-020-KAJ, 2004 WL 609331, at *6 (D. Del. Mar. 22, 2004) ("Reliability of a test is a necessary condition for validity") (citation omitted); PX 21 at p. 32. Dr. Gebhardt notably reports no reliability coefficients for incumbents' responses on this subject, although they had represented that that they would. DX 11 at ERN000386. As a result, the reliability of the "data" generated is unknown. PX 21 at p. 32. Reliability, however, is a necessary condition both for validity under the Uniform Guidelines and for admissibility under Fed. R. Evid. 702. "Unless a test measures individuals consistently, it can never been a valid test. Ramona L. Paetzold and Steve L. Willborn, The Statistics of Discrimination: Using Statistical Evidence in Discrimination Cases § 5:12. Survey responses like these, and others, collected by Dr. Gebhardt, lacking any measured or demonstrated validity, have no probative value.

36. It is also noteworthy how Dr. Gebhardt's own observations did not cause her to question her survey results—or even to measure and report their reliability. As already mentioned, as part of her job analysis, Dr. Gebhardt accompanied paramedics on 19 calls. Eleven of those calls involved patients who were not lifted because they were ambulatory—they walked to the ambulance on their own. Another involved a deceased patient who was not transported. Only one patient weighed more than 250 pounds. DX 37A at ERN000667-677. This paints a very different picture of the job of a CFD paramedic than the survey responses did, with their reliance on retrospective guesstimates rather than hard data.

D. <u>"Linking" Tasks to Abilities.</u>

37. After narrowing the task inventory from 156 tasks to 24 "essential" *physical* tasks, cutting out sizeable and important parts of the job, Dr. Gebhardt asked a small group, of 18 EMS personnel, to "link" each of those remaining 24 physical tasks to the physical abilities they

deemed needed to carry them out. For this purpose, Dr. Gebhardt gave 18 incumbents a questionnaire listing eight physical abilities, which they identified and defined as follows:

> STATIC STRENGTH – This is the ability to use muscle force for a single task lasting less than one minute which involves lifting, pushing, pulling or holding objects.
>
> MUSCULAR ENDURANCE – This is the ability to use muscle force for single or multiple tasks involving lifting, pushing, pulling, holding, or carrying which lasts two minutes or more. This ability also involves supporting or moving one's own body weight. It represents muscular endurance and emphasizes the resistance of muscles to fatigue.
>
> EXPLOSIVE STRENGTH – This is the ability to use bursts of muscle force lasting one second or less involving rapid movements of body parts to project an object or move one's own body.
>
> TRUNK STRENGTH – This ability involves the use of stomach and/or back muscles to perform a single task or repeated tasks. The ability involves the degree to which the trunk muscles do not fatigue when they are put under repeated or continuous strain.
>
> AEROBIC CAPACITY – This ability involves an increased demand on the lungs and circulatory (blood) system for tasks performed continuously for five minutes or more. This is the ability to perform tasks efficiently over long time periods without having to stop to catch your breath.
>
> EQUILIBRIUM – The ability to keep or regain one's balance, or to stay upright when in an unstable position. This ability includes being able to maintain one's balance when changing direction while moving or when standing motionless.
>
> FLEXIBILITY – This is the ability to bend, stretch, twist, or reach out with the body, arms, or legs.
>
> ANAEROBIC POWER – The ability to exert all-out physical effort to perform work or exercise tasks for brief periods of between 5 to 90 seconds. This ability involves repeated movement of the arms and/or legs for up to 90 seconds.

DX 37 (Appendix O). To complete this questionnaire, the 18 incumbents were also asked to rate the amount of each of these eight abilities required to carry out the tasks, on a scale from 1 to 7.

38. On the basis of the incumbents' ratings, Dr. Gebhardt concluded that CFD paramedics need seven abilities to perform their 24 most "essential" physical tasks—static strength, dynamic strength, explosive strength, trunk strength, equilibrium, flexibility and anaerobic power. She also concluded that they do not need aerobic capacity. DX 37 at ERN000593. Among the seven physical abilities deemed needed, the incumbents' rated four as more necessary than the others—trunk strength, static strength, explosive strength and flexibility. DX 37 at ERN000594. Among these, flexibility and equilibrium were critical to "maneuvering patients out of cramped spaces, performing CPR; administering patient care in a moving ambulance; [and] ascending/descending stairs with a patient on a stairchair, stretcher, or longboard; and lifting patients from floor ground level." DX 37 at ERN000593. However, no part of the final test battery tested applicants' flexibility or equilibrium.

39. The incumbents also quantified the amount of static strength and trunk strength they thought necessary to carry out job tasks, assigning static a mean rating of 4.84 (DX 37 at ERN000594) or roughly the strength needed to "lift a 50-lb box to shoulder height." PX 53 at ERN002925. For trunk strength they assigned a mean rating of 4.88 (DX 37 at ERN000594), equivalent to the strength needed to lift more than 50 pounds but less than 70 pounds (DX 37 at ERN002925).

40. Dr. Gebhardt conspicuously neglected to ask incumbents several significant questions, including four of the most critical questions that logically, necessarily, and historically have arisen when public employers attempt to validate a pre-employment physical performance test for a public safety position—namely, (1) Which of the physical abilities used on the job must applicants for the job possess on day one, upon entry, before training? (2) How much of each these physical abilities is required upon entry and how much can be gained after entry during

academy training? (3) Once on the job, what is the minimum level of each physical ability without which the job cannot be performed successfully? (4) To what extent, if any, does possessing more than the minimum required level of a physical ability correlate with better job performance? And (5) To what extent does the job require enough of an ability, after which having any more of that ability is merely redundant, making no discernable contribution to performance? The record in this case contains no answers to any of these questions with respect to the CFD paramedic position. And neither anecdotal evidence nor surmise can answer them. There must be systematic evidence based on a job analysis. There is none in this record. The failure to ask incumbents which physical abilities must be possessed on day one, before training is striking given that Dr. Gebhardt correctly asked exactly that question necessary with respect to personality traits. DX 37D at ERN000715-18. For physical "abilities," which she intended to make the focus of her test battery, she never asked that question.

41. Dr. Gebhardt's job analysis also raises a second unresolved and key concern, specifically, whether the incumbents who were asked to match tasks with physical abilities were qualified to do that. The terms used in the taxonomy of abilities that Dr. Gebhardt presented to the incumbents, terms such as "dynamic strength" and "anaerobic power," are not everyday terms. And the distinctions between some of them are abstract, including the difference between the three posited categories of strength (static, dynamic and explosive). The incumbents asked to apply this arcane taxonomy were not trained job analysts or physiologists, who could more reliably have been used to do the same linkage. The record contains no competent evidence that these incumbents understood these abstract categories in the same way that Dr. Gebhardt intended and interpreted them. One of the incumbents was apparently so unqualified that Dr. Gebhardt and HPS ignored his contributions. DX 37 at ERN000590.

IV. DEVELOPING THE TEST BATTERY, ATTEMPTING TO VALIDATE IT, AND SETTING A CUT SCORE

42. After completing the stages of the job analysis described above, there remained primarily three challenges for Dr. Gebhardt: (1) designing a test battery to measure the physical abilities that had been identified as needed for the job; (2) assessing whether performance on the test battery was "predictive of or significantly correlated with important elements of job performance," 29 C.F.R. 1607.5.B; and (3) setting a cutoff score to separate those who would be extended job offers from those who would not.

A. Picking Tests for Inclusion In the Final Test Battery.

43. To select tests for inclusion in the test battery, Dr. Gebhardt relied on their work on "previously developed tests," which is to say their familiarity with test batteries they had used in other work for other clients, often for positions other than the paramedic position, rather than anything specific to their job analysis of the paramedic position for the CFD. DX 37 at ERN000598; Tr. at 1413:14-20, 1414:3-9; 1414:14-23. Indeed, there is evidence that Dr. Gebhardt knew from the outset, even before she had secured a contract, which tests she intended to include in the final test battery. *See* DX 11 at ERN004393 (predicting likely use of "the Arm Endurance test").

44. On the basis of past experience, Dr. Gebhardt decided to consider the following ten "abilities" tests as candidates for inclusion in a physical performance test battery for Chicago: "Arm Endurance, Arm Lift, Leg Endurance, Leg Lift, Trunk Pull, Handgrip, Sit-Ups, Modified Stair Climb, Sit and Reach, and Stabilometer." DX 37 at ERN000598.

45. From the outset, Dr. Gebhardt knew, and advised the City, that use of these tests would have adverse impact against women. DX 11 at ERN004393; Tr. at 1668:24-1669:5.

46. The City had received an earlier proposal from another vendor promising its selection procedures, which also included a physical performance test, would "*not* result in adverse impact." PX 14(A) at ERN002842 (emphasis added). Dr. Gebhardt did not consult with exercise physiologists in conducting her job analysis or selecting the test battery for the PPT. By contrast, this vendor used two exercise physiologists, two physical therapists, six occupational therapists, and one vocational consultant to help with job analysis and test development. PX 14A at ERN002840.

47. There is no documentary evidence, including as Dr. Gebhardt admitted, no mention in her job analysis, that she, HPS, or the City ever considered or searched for alternatives to the PPT that she developed and the City implemented, which might have had less or no adverse impact. Tr. at 1697:10-12. DX 37.

48. The Uniform Guidelines state that:

> [W]henever a validity study is called for by these guidelines, the user should include, as a part of the validity study, an investigation of suitable alternative selection procedures and suitable alternative methods of using the selection procedure which have as little adverse impact as possible, to determine the appropriateness of using or validating them in accord with these guidelines."

29 C.F.R. § 1607.3B; *see also* 29 C.F.R. § 1607.16X; Questions and Answers To Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedure, 44 Fed. Reg. 11,996, 12,003 (Mar. 2, 1979), Question No. 48 ("Q. Do the Guidelines call for a user to consider and investigate alternative selection procedures when conducting a validity study? A: Yes . . ."). Dr. Gebhardt and the City violated this requirement of the Uniform Guidelines in this case.

B. The Need for a Job Performance (Criterion) Measure.

49. Having identified ten "abilities" tests as candidates for possible inclusion in the test battery, Dr. Gebhardt's next challenge was to "validate" one or more of them—that is, to generate evidence that might establish that a test battery including one or more of these "abilities" tests would be "predictive of or significantly correlated with important elements of job performance." 29 C.F.R. 1607.5(B)

50. The first step in any such validation effort necessarily required developing a measure of job performance. Simply put, test performance cannot be compared to job performance unless and until a stable and accurate measure of job performance exists. DX 37 at ERN000602.Validating any pre-employment test absolutely *requires* "having or being able to devise unbiased, reliable and relevant measures of job performance" or other comparably validated, reliable and accurate measures of "work behavior(s) or performance." 29 C.F.R. §§ 14B(2) and 16U; Tr. at 655:4-24.

51. Dr. Gebhardt, however, never had and never devised a reliable or relevant measure of job performance for the CFD paramedic position. Tr. at 668:13-19; 669:15-670:5; 1177:16-1178:12. Without such a measure she did not and could not successfully validate the PPT or adhere to the Uniform Guidelines.

52. In the employment testing context, this requisite measure of job performance is referred to as the "criterion"; a test or test battery used to predict it is referred to as a "predictor"; and, the extent, if any, to which scores on the test in fact predict or are significantly correlated with job performance is referred to as evidence of the test's "validity." Tr. at 1590:14-23, 1648:13-24.

53. By definition, the criterion validity of an employment test cannot be established without an accurate measure of actual job performance, leaving a hole in the City's evidence in this case.

54. "The most frequently used criterion measure [of job performance] is a supervisory rating of job performance." 1 B. Lindemann, P. Grossman & C.G. Weirich, Employment Discrimination Law 4-34 (5th ed. 2012).

55. Consistent with prevailing practice, Dr. Gebhardt considered using supervisory ratings of job performance as a criterion measure. To that end, she developed a supervisory rating instrument and collected supervisory and peer ratings of job performance for a small group of Chicago incumbent paramedics. DX 37 at ERN000603-04, ERN000635. These ratings assessed only incumbents' performance on physical tasks (rather than, for example, overall job performance, or the quality of patient care an incumbent provides). DX 37 at ERN000791-795.

56. The results of these performance ratings were illuminating. They indicated that women's job performance on physical tasks is very close to men's performance on those tasks. According to their supervisors and peers, the rated performance of the female CFD incumbent women on those tasks ranged from 90% to 93% of men's rated performance. DX 37 at ERN000637-639; Tr. at 1744:7-22. On that evidence, in any valid pre-employment test battery for CFD paramedic position, the mean test scores for men and women should be very close to each other: women's mean scores should be as close to men's as their job performance ratings are. Any significantly larger or smaller gap between them would, by definition, demonstrate that a test was not job-related.

57. The touchstone of job-relatedness is the close correlation between test performance and job performance. When the two are not strongly correlated, then a test is not

job-related. Based on the supervisory and peer ratings collected by Dr. Gebhardt and HPS, that is the case here: Dr. Gebhardt's own work indicates that women's job performance as measured by peer and supervisor ratings was much closer to their male counterparts than their PPT scores. Tr. at 1744:7-1745:1; PX 23 at p. 6. The peer and supervisor ratings indicated that women's physical abilities were 90-93% of men's physcial abilities. By contrast, over six administrations of the PPT to CFD paramedic applicants between 2000 and 2009, the average difference between men and women was much, much larger: women's test scores were, on average, about 300 points lower than men's, so that a female applicant in the highest 85% of the women tested had, on average, the same score as a man in the lowest 15% of all men tested. PX 21 at Table 2 and p. 19.

58. In place of supervisory/peer ratings, which showed only small differences between the relevant physical performance of incumbent CFD men and women, Dr. Gebhardt resolved to use work "sample" tests as a criterion measure instead. This meant using scores on one set of tests (the work "samples") to try to validate another set of *tests* (the PPT)—without ever bothering to measure *actual job performance*. Both conceptually and in its execution this strategy was a failure. Tr. at 477:11-25.

59. When the thing to be predicted (the criterion) is job performance, there is no substitute for, or legitimate way around, measuring actual job performance. The criterion measure of job performance can be, and frequently is, supervisor or peer ratings. In appropriate cases, it can also be another measure of worker productivity or behavior—like production counts, or absenteeism. *See* Tr. at 662:8-14. But to establish the validity of a pre-employment test, the criterion must be a measure of *actual* job performance or behavior, not scores on another test that has itself never been validated against job performance. Tr. at 654:18-22, 655:14-24,

665:12-15, 666:5-7, 666:18-25, 667:1-668:19, 669:5-17. Comparing performance on the work "sample" tests and performance on the PPT does not demonstrate a correlation between the PPT and job performance. It only shows a correlation between the PPT designed by Dr. Gebhardt and another set of tests designed by Dr. Gebhardt. Tr. at 477:13-16. This problem with validating physical tests against other tests was identified in the published literature many years ago. PX 21 at pp.34-35. It has also been rejected by courts. *See Guardians Ass'n of New York City Police Dep't, Inc. v. Civil Serv. Comm'n of City of New York*, 633 F.2d 232, 244 (2d Cir. 1980), *aff'd sub nom. Guardians Ass'n v. Civil Serv. Comm'n of City of New York*, 463 U.S. 582 (1983) (rejecting the "flawed argument" that job-relatedness can be found on the basis of even a high correlation between the results of two separate testing practices, neither of which by itself has been validated according to accepted methods).

60. The City would have the court find job-relatedness on the basis of a correlation between the results of two separate sets of tests (the work "sample" tests and the three "abilities" tests included in the final PPT battery), *neither* of which has ever been validated as a measure of job performance. The Court cannot accept this argument. Correlation to actual job performance remains unaccomplished. The City cannot, by substituting performance on the work "samples" as a correlate to performance on the PPT, avoid the obligation to show job relatedness—a correlation to actual job performance.

61. Even if it were possible in theory to conduct a validation study consistent with the Uniform Guidelines and Title VII by correlating one test with another without ever bothering to validate either one against a measure of actual job performance, which is doubtful, what Dr. Gebhardt, and the City did here falls very far short.

62. The strategy they employed was the following. First, Dr. Gebhardt devised three work "sample" tests, intending that measurement of incumbents' performance on these tests could substitute for measuring *actual* job performance. Second, Dr. Gebhardt, and the City recruited a group of 52 incumbents as a "validation sample," to take both the three work sample tests and each of the ten "abilities' tests being considered for inclusion in the physical performance test battery, hoping that the performance of these 52 incumbents on the work "sample" test would correlate with their performance on the physical performance battery (i.e., with both high and lower scores on the work sample tests scoring correspondingly high or low on the physical performance battery). Third, they reasoned that if their work "sample" tests were accurate enough miniature replicas of actual job performance, and if the correlation between scores on the work "sample" test and the physical performance battery were strong enough, then: (a) the existence of a correlation should imply that successful test performance predicts success in actual job performance; and, (b) the inference of success in actual job performance should be generalizable to the larger population of all CFD incumbents and/or all applicants for CFD paramedic positions. This reasoning was deeply flawed because the work "sample" tests were not accurate samples of actual job performance and because test results based on the performance of the 52 incumbents were not generalizable.

C. <u>The "Work Sample" Tests Were Not Accurate Replicas of Actual Job Performance.</u>

63. Dr. Gebhardt devised three work "sample" tests. They were not standard work "samples." The first, which she referred to as the "lift and carry" test, timed each of the 52 incumbents in the validation sample as they lifted a piece of equipment, weighing from 17 to 29 pounds, carried the equipment 35 feet to a set of stairs, climbed up the stairs (19 steps), placed the equipment on the landing, lifted another piece of equipment, descended the stairs, and placed

the equipment in the starting area. The scoring for this test was the time to complete five cycles up and down the stairs. DX 37N at ERN000781.

64. The second work "sample," which Dr. Gebhardt referred to as the "stairchair push," involved removing a piece of equipment from an ambulance, pushing a simulated patient-loaded stairchair up a ramp, down a ramp, through a 60-foot obstacle course, returning the stairchair to the start area, and then returning the equipment to the ambulance, while wearing a quick response bag. The scoring for this test was the time to complete four cycles. DX 37 at ERN000602; DX 37(N) at ERN000784; Tr. at 1752:10-25.

65. The third work sample test, referred to as the "stretcher lift," required each of the 52 incumbents in the validation sample to lift one end of a stretcher to an arm locked height, hold it for 20 seconds, lower the stretcher and rest for five seconds, lift the stretcher again, this time to a height of 34.5 inches, hold it at that height for five seconds, and then lower the stretcher and rest for 30 seconds. For the first cycle, the stretcher apparatus weighed 90 pounds. After each cycle was successfully completed, ten more pounds were added. The incumbents were instructed to continue lifting until they could no longer complete the cycle or had completed 13 cycles, whichever occurred first. This test was scored by the number of cycles completed and the weight of an incumbent's last successful lift. DX 37 at ERN000603.

66. For at least two sets of reasons, these work sample tests were neither measures of nor faithful simulations of actual job performance. First, both the "'lift and carry" and the "stair chair push" were "speeded" events. They were timed, with scoring based solely on speed. Faster times were better scores. The problem with this is clear. Dr. Gebhardt's job analysis presents no data about how quickly paramedics either do or should be able to perform these tasks or tasks like them and no justification for having a measure of job performance that depends solely on

how quickly a paramedic performs any task. DX 37 at ERN000602-603; Tr. at 1751:24-1752:25. Even if one could agree, based on an adequate job analysis, that the "lift and carry" and "stair chair push" accurately simulated actual job behaviors, which they do not, there is no basis in the evidence for concluding that the most successful paramedics will be the ones who perform these tasks in the shortest time. *See* PX 22 at pp. 29-30. Dr. Gebhardt's job analysis and validation report provide no foundation for evaluating or scoring the work sample tests on the basis of speed. There is no empirical data or measure of necessary "speed" on Dr. Gebhardt's list of 156 paramedic tasks, much less on her narrowed list of 24 "essential" "physical" tasks. DX 37C at ERN000681-689; DX 37K at ERN756-757.[9] In prior studies for other jobs, Dr. Gebhardt had attempted to quantify the speed of movement necessary, undertaking pace and tempo studies. Tr. at 1747:19-1748:4. She made no such effort here. Her job analysis did not ask or determine how quickly which tasks should be performed. Tr. at 1749:22-1752:6.

67. Second, neither the "lift and carry," nor the "stair chair push," nor the "stretcher lift" were accurate simulations of actual job behaviors. PX 21 at p. 36. Standard work "samples" replicate the job. By contrast, these did not. They did not replicate either the manner, method or means by which CFD paramedics perform their jobs. By Dr. Gebhardt's own admission, the "stretcher lift," for example, does not sample any actual work behavior. Tr. at 1754:23-24. It was

[9] E.g., *Berkman v. City of New York*, 536 F. Supp. 177, 215-16 (E.D.N.Y. 1982) *aff'd*, 705 F.2d 584 (2d Cir. 1983) ("Even if one could agree based on an adequate job analysis that the five job sample tasks simulate actual job behaviors, there is . . . an assumption . . . which has no basis in the evidence or common sense, that is, that the most successful firefighter will be the one who performs these five tasks in the shortest time . . . . [N]othing in the [validation] study shows that performance of the five tasks . . . , assuming them to be important or critical work behaviors, is to be rated simply according to the speed with which they can be performed seriatim. Accordingly, I can find no basis in the Maryland study for determining that Exam 3040 is criterion valid").

a one-person lift, while the actual job involves two people lifting a stretcher.[10] Further, the stretcher lift test involved and measured timed, repeated lifts, in quick succession, each successively heavier than the last, as if paramedics were called on to lift patients seriatim, one after another in a short span, with each patient heavier than the last until the point of fatigue or the thirteenth and heaviest patient. That scenario bears no resemblance to any of the 156 tasks identified by Dr. Gebhardt and HPS in their job analysis of the paramedic job. See DX 37C; PX 22 at pp. 29-30. It was not job related. It was not predictive of paramedic job performance. Tr. 665:7-15, 667:24-668:6.

68. Similarly, the "stairchair push" also was not a work "sample." It did not replicate actual job behaviors. Neither the job analysis nor the trial testimony provide any evidence that paramedics do, or should, push a stairchair at maximum speed, or even that pushing as quickly as they can would be safe. In addition, the obstacle course used for the stairchair push required pushing a stairchair up a ramp, on the misunderstanding that that simulates what typically occurs when patients arrive by ambulance at the hospital. But, in point of fact, the relevant EMS System Policies and Procedures provide that patients may not be transported to the hospital in a stairchair. PX 21 at p. 36. The "stairchair push" was not job related. It was not predictive of paramedic job performance. Tr. 665:16-666:7, 667:24-668:6; PX 22 at pp. 29-30.

69. Likewise, the "lift and carry" work sample, although bearing a closer resemblance to actual job behaviors, also did not faithfully replicate actual work behavior. It called on incumbents to climb and descend stairs at maximum speed (since faster times were better scores), despite the fact that nothing in the job analysis or the record provides any evidence that

[10] Only two of the twenty-four physical tasks that Dr. Gebhardt isolated as "essential" were performed without assistance. DX 37K at ERN00756 (Tasks 11 and 12). By contrast, all of her work "samples" were to be performed by one incumbent, without assistance.

paramedics do, or should, ever take stairs at maximum speed (or even at what speed they do or should climb stairs). Similarly, nothing in the job analysis or record suggests that paramedics ever climb two flights of stairs only to immediately turn around and descend the same stairs, followed by a repetition of the same behavior three more times. Among the several paramedic witnesses who testified at trial, none established, or even attempted to establish, any resemblance between this work "sample" and actual work behaviors. Finally, Dr. Gebhardt reported a very low reliability, of .503, for the "lift and carry" work, lower than for any of the other "work samples," making reliance on it as a basis for physical test selection or validation questionable at best. DX 37 at ERN000626. The "lift and carry" was not job related. It was not predictive of paramedic job performance. Tr. 666:8-25, 667:24-668:6.

D. <u>The 52 Incumbents In the Validation Sample Were a Non-Random, Unrepresentative Sample, Whose Test Results Are Not Generalizable.</u>

70. The idea of forming a validation sample of CFD incumbents who would take both the "work sample" tests and the physical performance battery was absolutely crucial to Dr. Gebhardt's attempt to validate the physical performance battery. It was this group of CFD incumbents whose performance on the work "sample" tests would, she hoped, stand in for a criterion measure of actual job performance; whose performance on the work sample tests would also, she hoped, correlate with their performance on the physical performance battery; and whose performance on both sets of tests would, she hoped, be generalizable, allowing the inferential leap that if *applicants* performed well on the physical performance battery they could be presumed to perform well on the job—and, conversely, that applicants who did not score well on her tests would be "unsuccessful" job performers.

71. This chain of inferences was tenuous. But it became untenable given the composition of the validation sample of CFD incumbents.

72. An analysis is only as good as the data on which it rests and "[i]nferences from the part to the whole are justified only when the sample is representative." Reference Guide on Statistics, in Reference Manual on Sci. Evid. 211, 217 (Fed. Jud. Center) (3d ed. 2011). These conditions were violated here, rendering Dr. Gebhardt's use and reliance on the validation sample unreliable, both as a matter of fact and law. Tr. at 657:21-659:4.

73. Initially, Dr. Gebhardt told the City that she needed a validation sample comprised of 125-150 incumbent paramedics in order to establish a valid criterion for job performance and validate a physical performance battery. PX 54 at ERN002893, PX 19A at ERN004916. Subsequently, HPS revised that number downward slightly, offering that "approximately 110 incumbents would suffice. PX 19(B) at ERN002581. In response, the City selected particular paramedics whom it invited to volunteer. PX 62; PX 19(C). In the end, CFD was only had 52 CFD incumbents to participate in the validation sample, less than half the number originally stated as needed. DX 37 at ERN00607.

74. Of the 52 incumbents who ultimately volunteered to participate in the validation sample, only 25 were working paramedics. Among the remaining 27, nine were firefighter-paramedics, fresh on the heels of completing their training at the firefighter Academy, with its rigorous physical regimen; 17 were paramedics in charge; and, one was an ambulance commander. DX 37 at ERN000607, ERN000609; PX 62 at ERN003405; Tr. at 1722:9-12.

75. The formation and composition of the group did not comport with any scientific sampling standard. The group was not randomly selected. Tr. at 1721:2-10. It was a "convenience" sample, made up of self-selecting rather than randomly selected volunteers, assembled haphazardly, and neither comparable to nor representative of the larger overall

population of CFD paramedics.[11] No adjustments were made to take the effects of volunteerism into account. In addition, and importantly, this was a much more physically fit group than CFD paramedics generally (a foreseeable result for a group volunteering to submit to physical performance testing). The CFD women included in this group were especially atypical: as Dr. Gebhardt admitted in an early version of her validity report, they "performed better than all or most of the women in other physical performance validation projects." PX 121 at ERN004834. On the arm endurance test, which they were given as part of the physical performance battery, "the Chicago Paramedic women's validation sample performed better than all other comparison (12) validation projects." *Id*. Dr. Gebhardt excised this admission in her final validation report. DX 37. She eliminated any discussion of it, concealing these facts.

76. The men in the "validation" sample were also atypical and unrepresentative of the population of CFD paramedics overall. On the leg "lift," which was administered to them as part of the physical performance battery, their mean score was 281.9. DX 37 at ERN000611. That is significantly better than the mean for the male applicants who ultimately took the PPT (98% of whom passed and went on to become CFD paramedics): the mean for male applicants was 245.11. PX 21 at p. 20. On the arm "endurance" test, the mean score of CFD male incumbents was 243.5, while for male applicants the mean was 217.96. DX 37 at ERN000611; PX 21.

77. Among the male incumbents in the validation sample, the nine firefighter/paramedics were especially unrepresentative. Their mean test performance was significantly better than the mean for the rest of the validation sample, even as the mean for the

[11] *See DeKoven v. Plaza Assocs.*, 599 F.3d 578, 581 (7th Cir. 2010) ("The sample drawn by the . . . expert is what is called a 'convenience' sample—convenient to the sampler—as distinct from a 'representative' sample—representative of the population sampled.").

validation sample was significantly higher than for the population of CFD paramedic generally. DX 37 at ERN000620-21.

78. This highly fit group of 52 Chicago incumbents resembled the "straight-A" gym students in a class. Tr. at 670:6-8. And for the same reason that one does not set the curve the curve on a classroom test or hazard predictions about the performance of the rest of the class, based only on the grades of atypical students who have excelled on the particular, narrow test being given, so, too, it was unscientific and unreliable for Dr. Gebhardt to treat the performance of this "validation" sample of 52 Chicago incumbents as any basis for prediction. The performance levels of the 52 Chicago incumbents provided an inappropriate basis for setting a passing score for the PPT, for distinguishing between satisfactory and unsatisfactory job performance, or for predicting the job or test performance of other CFD incumbents or applicants for CFD paramedic positions.

79. The fact that this validation sample group was neither randomly drawn from the population of CFD incumbent paramedics nor representative of the population of CFD paramedics (or the labor pool) introduced incurable error into the attempt to validate the PPT. It renders Dr. Gebhardt's use and reliance on data drawn from this sample statistically unreliable and inconsistent with accepted statistical methods and principles and Fed. R. Evid. 702. Tr. at 490:12-23 and 657:8-659:2; PX 21 at pp. 42-43.

80. With a sample of only 52, which was far smaller than she had wanted, and a restricted range of scores within that validation sample (because the range of scores of the 52 CFD incumbents was compressed as a result of their exceptional performance), Dr. Gebhardt resorted to a kludge. She decided to "transport" "supplemental" data from New York City, in order to "expand" both the sample size and the range of scores. DX 37 at ERN000655; Tr. at

1737:25-1738:15. For this purpose, she added data pertaining to 87 paramedics from New York City to her Chicago validation sample. But that only compounded problems, further polluting the existing validation sample. The New Yorkers not only were themselves a non-random sample, they were a non-random subset of a larger non-random sample of New York City paramedics. Tr. at 1739:13-1741:11. Dr. Gebhardt did not "transport" a full data set from New York. Rather, she selected only a portion of it. She made no adjustments to take the effects of this cherry picking, nor for the effects of adding more "volunteers" to the sample, and no analysis to evaluate whether the physical performance of the New Yorkers was representative either of paramedics in their own department or of CFD paramedics.

81. The Uniform Guidelines provide that "the sample subjects . . . should be representative of the candidates normally available in the relevant labor market," and that where "samples are combined . . . , attention should be given to see that such samples are comparable in terms of the actual job they perform and other relevant factors likely to affect validity." 29 C.F.R. § 1607.14B(4). In disregard of this requirement, Dr. Gebhardt's job analysis simply does not provide such details. Indeed, the job analysis does not even disclose that her "supplemental" sample was drawn from New York. It merely states that supplemental data were "collected in the same type of setting as the Chicago EMS validation study" and that these supplemental data points were drawn from "incumbent EMS personnel" with "similar" "backgrounds, (i.e., EMS)." These conclusory statements do not meet the standards of the Uniform Guidelines for providing details about samples used. *See* 29 C.F.R. § 1607.15B(6). Importantly, Dr. Gebhardt also does not mention how the New York incumbents were selected for testing in New York or provide any basis for evaluating whether they were representative of pool of New York City paramedics generally or of any other group.

E. Final Selection of Component Tests For the PPT.

82. After testing the 52 CFD incumbents in the validation sample on her three work "sample" tests and the ten "abilities" tests under consideration for inclusion in the physical performance test battery, Dr. Gebhardt examined various combinations of their scores on these measures for correlations. She concluded that the evidence suggested that only three of the ten possible "abilities" tests under consideration should be included in the final physical performance test battery. The result was a final battery consisting of the Arm Endurance, Leg Lift, and Modified Stair Climb tests. Dr. Gebhardt made this decision despite the fact that, of all the tests she was considering for inclusion in the physical performance test battery, the "Arm Endurance" and "Leg Lift" tests had produced among the greatest disparities between the genders. DX 37 at ERN000611-12; PX 21 at pp. 9, 30-31. She rejected several available "abilities" tests with smaller gender differences, like the "Sit and Reach Test" (which measured flexibility) and the "Equilibrium Test" (which measured balance).

83. On the "Sit and Reach" and "Equilibrium Tests" in particular, the data showed that women scored *higher* than men. DX 37 at ERN000611-12; PX 21 at pp. 30-31. The job analysis also showed that both flexibility and equilibrium were very important to successful job performance. In fact, the job analysis showed that flexibility was rated the third most important ability for paramedics, surpassed only by trunk strength (which the PPT does not measure) and static strength. The final test battery did not purport to measure flexibility or equilibrium, despite incumbents' high ratings of both in their assessment of the physical abilities necessary to perform the job. DX 37 at ERN594.

F. The Tests Selected: The Arm Endurance Test.

84. The "Arm Endurance" test involved kneeling and pedaling an arm crank ergometer against 50 watts of resistance continuously for two minutes. The goal is to complete as

many pedal revolutions as possible, with a candidate's score equaling the number of pedal revolutions completed. This test primarily measures upper body anaerobic output when cranking your arms. Tr. at 444:24-445:9. Dr. Gebhardt's name for the test ("Arm Endurance") reveals a fundamental misunderstanding: this test does not measure endurance or muscular endurance in the arms. Nor does it require a high level of cardiovascular fitness or aerobic energy transfer. It measures short-term power, anaerobic power capacity, the ability to perform very short duration effort, not endurance. Tr. at 446:6-15, 467:6-20. It is not job-related for the position of paramedic. Tr. at 516:4-10. *See also* Tr. at 427:8-13, 445:25-445:5, 1178:13-20; PX 23 at p. 9; Tr. at, 810:14-811:19, 1039:17-20, 1249:18-22.

85. Dr. Gebhardt's job analysis reports a correlation of .64 between incumbents' performance on the "Arm Endurance" test and their performance on her work sample tests. DX 37 at ERN000648. A correlation at that level, even if it were comparing relevant variables and was reliable (with neither of those conditions satisfied here), would account for only 41% of the variance between that test and incumbents' performance on the work samples. *See* Tr. at 478:1-9, 1708:4-1709:2 (explaining that squaring a correlation coefficient gives the proportion of variance). Even if the work samples were valid criterion measures of actual job performance, which they are not, less than half the variation in work sample performance would potentially be explained by performance on the "Arm Endurance" test, providing only a weak, if any, foundation for using performance on the arm crank ergometer as any basis for predicting performance on the work "sample" tests, let alone any foundation for using an arm crank ergometer to predict performance on any paramedic job task. *See* Tr. at 477:6-478:9.

G. <u>The Tests Selected: The "Leg Lift" Test.</u>

86. The "Leg Lift" test involved standing with knees bent at a fixed angle and back straight, while grasping an immovable bar connected to a chain between your legs, and then

exerting upward force against the bar by attempting to extend your legs. PX 4 at ERN000270. The goal was to apply as much upward force as possible, and scoring was equal to pounds of force applied. This is a test of "static" or "isometric" strength which measures a very specific construct: leg force generated by a select group of muscles in our leg, with the knee bent at a fixed angle, with these results highly specific and sensitive to the particular angle. Tr. at 435:3-436:9, 439:19-22, 441:12-14, 442:5-9, 457:9-14, 505:22-506:1.

87. Dr. Gebhardt's name for this test ("leg lift") evidences a fundamental misconception: this test involves no movement and no lifting. It involves exerting static force against an object that is not lifted and does not move. Tr. at 441:1-5. For that reason, it is unlike anything paramedics ever do on the job. There is no evidence that paramedic job performance *ever* involves a static or isometric lift. When they exert force it is to move things, such as stretchers, patients, equipment. Tr. at 440:1-18; 441:6-17.

88. This test was not job related; it did not measure, predict or replicate any of the 24 tasks Dr. Gebhardt identified as essential to the job or any other task a paramedic performs as part of the paramedic job. Tr. at 516:4-10. *See also* Tr. at 427:8-13; 438:6-17; 439:11-13; 1687:16-1689:1, 1178:24-1179:12; PX 23. *See also* Tr. At 812:3-8, 814:3-14, 1040:1-6, 1078:2-14; *see also* Tr. at 417:16-418:15.

89. Dr. Gebhardt never measured any incumbent's (or applicant's) performance of any of those 24 "essential" physical tasks. And having never measured performance of any of those tasks, she also, of course, could not and did not correlate PPT performance to performance of any of those tasks. The inferential leaps the City asks the Court to make are considerable (and given the flaws both in test design and in Dr. Gebhardt's statistical methods also impossible).

90. In addition, men respond to this test better than women; many women find the position required to perform the test awkward and even uncomfortable. Tr. at 269:24-270:3, 814:3-14; s*ee also* Tr. at 436:10-437:8, 443:4-14. In addition, for more women than men, learning to succeed at this test involves a steep learning curve, such that without pretest training or practice on this item women are at a disadvantage. Tr. at 436:17-23; PX 23 at p. 10. However, proper training improves performance on this test, and affording candidates a greater opportunity to familiarize themselves with the specific testing apparatus would have reduced that disadvantage and reduced disparate impact. Tr. at 437:7-8, 443:7-25, 453:14-16; PX 23 at p. 10.

91. Dr. Gebhardt's job analysis reports a correlation of .68 between incumbents' performance on the "Leg Lift" test and their performance on her "work sample tests. DX 37 at ERN000648. A correlation at that level, even if it were comparing relevant variables and was reliable (with neither of those conditions satisfied here) would potentially account for only 46% of the variance between that test and incumbents' performance on the work samples. *See* Tr. at 1708:4-1709:2 (explaining that squaring a correlation coefficient gives the proportion of variance). Even if the work "samples" had been valid criterion measures of actual job performance, which they are not, less than half the variation in work sample performance would potentially be explained by performance on the "Leg Lift" test, providing only a weak, if any, foundation for using performance on the "Leg Lift" as any basis for predicting performance on the work "samples," let alone as a basis for predicting performance on any paramedic job task. *See* Tr. at 477:6-478:9.

H. <u>The Tests Selected: The Modified Stair Climb (Or Step Cross).</u>

92. The "Modified Stair Climb" test involved walking up a two-stair platform, carrying an 18-pound bag in one hand and a 19-pound bag in the other, crossing the top of the platform, descending two stairs on the opposite side, then pivoting or rotating and crossing again.

The goal was to complete as many crossing as possible in four minutes, with a candidate's score equaling the number of crossing completed. PX 4 at ERN000269.

93. Dr. Gebhardt's belief that this test measures "muscular endurance of the muscles in your upper and lower extremities" (PX 4 at ERN000269) evidences a fundamental misunderstanding. The test measures cardiovascular fitness. It does not measure muscular strength or endurance. Tr. at 450:25-451:7, 465:21-22; PX 23 at pp.8-9. It does not resemble step tests that have been validated in the scientific literature as measures of cardiovascular fitness. Tr. at 451:19-452:22.

94. This test also requires rotating or pivoting, balance, and a requirement that both feet must touch the floor and the end of each descent. PX 4 at ERN000269. Those are confounding factors that will have a material effect on many candidates' scores, while also requiring subjective judgments by test proctors. *See* Tr. at 451:19-452:18, Tr. at 330:1-331:13, 1337:5-14, 1338:7-1339:1, 1370:17-23, 1377:7-1378:1.The test is not job-related for the position of paramedic; it does not measure, predict, or replicate any of the 24 tasks Dr. Gebhardt identified as essential to the job or any other any task a paramedic performs as part of the paramedic job. Tr. at 314:19-315:3, 516:4-10, 812:3-13, 1039:7-1040:6. *See also* Tr. at 427:8-13, 451:8-18.

95. Dr. Gebhardt's job analysis reports a correlation of .56 between incumbents' performance on the "Modified Stair Climb" test and their performance on her "work sample tests (DX 37 at ERN000648), which she unjustifiably relied on as a proxy for the requisite criterion measure of actual job performance, which she also did not have. A correlation at that level, even if it were comparing relevant variables and was reliable (with neither of those conditions satisfied here) would potentially account for only 42% of the variance between that test and

incumbents' performance on the work samples. *See* Tr. at 1708:4-1709:2 (explaining that squaring a correlation coefficient gives the proportion of variance). Even if the work "samples" were valid criterion measures of actual job performance, which they are not, less than one third of the variation in work "sample" performance would potentially be explained by performance on the "Modified Stair Climb" test, providing only a weak, if any, foundation for using performance on the work "samples," let alone for predicting performance on any actual paramedic job task. See Tr. at 477:6-478:9.

96. Dr. Gebhardt contended that if the "Arm Endurance," "Leg Lift," and "Modified Stair Climb" tests are considered together they might cumulatively account for approximately 64% of the variation in work "sample" performance. Tr. at 1821:17-1822:11. That would leave more than a third unaccounted for, a very modest basis for prediction, almost entirely an artefact of correlating one physical test with another rather than with job performance (Tr. at 477:13-25), and insufficient in view of the severity of the disparate impact here.

97. Fundamentally, the decision to place the "arm endurance," "leg lift," and "modified stair climb" tests in a battery to be used to select paramedic hires was ill-advised because the job-relevant skills they appeared to be intended to measure (but didn't), could have been measured directly, also resulting in more accurate measurements. Tr. at 453:17-454:7.

98. The ability to lift someone on a stretcher can be measured by having applicants lift a stretcher. Similarly, the ability to go up and down stairs carrying equipment can be measured by having someone go up and down actual flights of stairs, rather than crossing a platform with two steps up and two steps down and then turning and pivoting and crossing again. It's not difficult to simulate these tasks. *Id.* at 478:10-16.

99. There was no good reason to complicate things with indirect measures, using contraptions like an arm ergometer or a leg dynamometer, when performance could have been measured so much more directly and without having to worry about and make adjustments for the "slop" and other variables that were needlessly injected to the process of measurement, exacerbating the adverse impact on women. This is self-evident not only conceptually but also from the point of view of physiology. Tr. at 453:17-454:7, 478:10-16.

100. The paramedic job does not require paramedics to pedal an arm crank ergometer; walk across a step platform, pivot and walk back; or hold an immovable bar between their legs while looking upward and trying to exert pressure upward. None of that is expected on the job. The contraptions used for the PPT do not replicate resemble motions or movements actually used on the job. They also do not measure or predict the ability to perform motions or movements used on the job. Tr. at 427:8-10, 438:6-439:2, 439:11-13, 445:25-446:2, 451:8-10.

I. The Reliability of the PPT is Unknown.

101. The "reliability" of a test refers to the extent to which its scores are free from random error. PX 21 at p. 32. Reliability can be measured in different ways, which focus on different sources of random error. One important measure of reliability, known as "test-retest" reliability, refers to the extent to which, if a test were given a second time, the same people would score at the top, bottom and middle of the score distribution—or, equivalently, the likelihood that candidates scoring within a given range below the 935 cut-off score might have succeeded on a re-administration of the same test, after a brief rest period.

102. The test-retest reliability of the PPT is unknown; no test-retest data was gathered. *See* DX 37. Without such data, it is difficult, if not impossible, to determine whether test scores on this test provide a stable measure of anything.

103. The failure to gather test-retest data on the PPT stands in stark contrast to Dr. Gebhardt's effort to gather precisely such data for her work "sample" tests. *See, e.g.*, DX 37 at ERN000625-26. In the contract, HPS recognized this, stating that it would provide this data. *See* DX 14 at ERN004240 (The reliabilities of the tests and the criterion measures will be assessed using standard procedures (e.g. test-retest reliability, intertrial reliability, interrater agreement.

104. The absence of an accurate, or any, measure of the reliability of the PPT is dispositive of the City's attempt to establishing the validity of the PPT: without evidence establishing reliability, a finding of validity is not possible. PX 21 at pp. 32-33.

J. <u>The Cutoff Score.</u>

105. Even the most meticulous job analysis and utmost care in test construction cannot save a test if the cut score used to distinguish passing from failing scores is not independently job related. It is the employer's burden to prove that "all aspects of the test including the method of scoring it" are job related. *Evans v. City of* Evanston, 881 F.2d 382, 384 (7th Cir. 1989). A cut off score unrelated to job performance that results in adverse impact violates Title VII.

106. The formula Dr. Gebhardt and HPS devised to score the PPT was:

(7 * Modified Stair Climb) + (1 * Leg Lift) + (2 * Arm Endurance)

DX 37 at ERN000662. When this sum equaled or exceeded 935, an applicant was deemed to have passed the PPT. Scores lower than 935 were deemed flunking scores. *Id.*

107. The City introduced no evidence correlating a score of 935 to any minimum level of ability measured by any of the three parts of the PPT: the "Arm Endurance" test, the "Leg Lift," or the "Modified Stair Climb." It could not. Given the formula for scoring the PPT, which took a weighted sum applicants' scores on all three measures, it is undeniable that a score of 935 cannot insure or correspond to any minimum level of performance on any of those three tests.

108. It is also undeniable, given Dr. Gebhardt's failure to identify or use a measure of actual job performance, that the City has not demonstrated any correspondence between a score of 935 and any level of job performance. The City introduced no evidence correlating scores on the "Arm Endurance" Test, the "Leg lift," or the "Modified Stair Climb" with any minimum level of ability needed to perform any of the 24 essential tasks identified as a result of the job analysis. There is no evidence correlating performance on any component of the final PPT battery with performance of any task identified in the job analysis.

109. Inherently, any scoring system that functions as Dr. Gebhardt's system for scoring the PPT did, by summing weighted scores from multiple sub-tests, "allows for an individual to compensate for a low score on one test with a higher score on another test." DX 37 at ERN000657. By definition, that makes it possible to pass without "attain[ing] a[ny] specific score on" any one of the tests in the test battery. *Id.*

110. As a result, any "compensatory" model for scoring necessarily implies: (a) that none of the abilities being tested for are essential to job performance, since otherwise it would be unacceptable to score the test in a manner that allows a superior performance on one to compensate for poor performance on another; and, (b) if they are essential, then the passing score does not represent or predict a minimally acceptable level of job performance because it does not insure any minimum level of ability on any of the measurements being taken. *See* Tr. at 470:3-471:10; PX 23 at pp. 16-17. These facts independently compel the conclusion that PPT's cutoff score of 935 was not job related.

111. However, the effort to establish a defensible cut off score on the PPT was also futile from the outset. Given the invalidity of each of the three tests that make up the PPT, which do not predict and are not significantly correlated with actual job performance, there is no

passing score methodology that could have salvaged the PPT. PX 22 at p. 38. These facts compel the conclusion that neither the PPT nor the cut scores were job related. Even so, as discussed below, the manner in which the cutoff score of 935 was set provides myriad additional proofs that that cutoff was not job related.

112. The passing score for the PPT was set in three steps. First, Dr. Gebhardt ranked all 137 Chicago and New York incumbents in her validation sample by deciles (ten percent intervals), according to her scores on the three work "sample" tests—i.e., the "Lift and Carry," "Stretcher Lift," "Stair Chair Push." DX 37 at ERN000657. This step baked two spurious judgments into every subsequent step in this model, namely: (1) that the work samples replicated actual job behaviors, which they did not, and (2) that work sample scores could substitute for "criterion" measures of *actual* job performance, which they cannot. These judgments introduced incurable error into the setting of the cutoff score, pre-ordaining that any cutoff based on these judgments could not be job related or consistent with business necessity.

113. Next, Dr. Gebhardt acknowledged that professional standards require setting a cutoff score for a pre-employment test require setting a "physical test cutoff score[ ] . . . [at] the minimum acceptable level." DX 37 at ERN000660; Tr. at 1732:1-5.

114. By definition, establishing a cutoff at the minimum acceptable level needed for successful job performance, as required by professional standards, presumes knowledge of what that minimum acceptable level of job performance is. There simply can be no assessment or measurement of the strength of the relationship, if any, between test scores and job performance without a measurable and reliable index or measure of job performance. This posed an insuperable obstacle for Dr. Gebhardt. In this situation, she punted. They decided to establish it by *ipse dixit.* They simply postulated that a minimally acceptable job performance level would

be a score 1.21 standard deviations below "the mean for the Chicago validation combined criterion measure (∑ 3 work samples)"—meaning 1.21 standard deviations below the mean of Chicago incumbents' composite scores on the stairchair push, lift and carry, and stretcher lift. DX 37 at ERN000660-661. This decision introduced yet another source of incurable error into the setting of the cutoff score and, once again, pre-ordained that any cutoff set by this method could not be job related or consistent with business necessity.

115. There is no evidence that any individuals in the Chicago validation sample were in fact unsatisfactory job performers. Indeed, as discussed above, Dr. Gebhardt neither possessed nor ever devised any measure of incumbents' actual job performance—except supervisor and peer ratings which, however, she refused to rely on after collecting them. Tr. at 668:13-19. In setting the "minimum acceptable level" at 1.21 standard deviations below the mean, Dr. Gebhardt simply decided by fiat that 15.3% (or roughly one-sixth) of the Chicago validation sample would have to be labelled "unsatisfactory" performers—regardless of their actual level of job performance. In point of fact, and proving the disconnect between Dr. Gebhardt's performance on the PPT and Dr. Gebhardt's arbitrary definition of "unsatisfactory" job performance, of the six CFD incumbent paramedics who fell below her 1.21 standard deviation cutoff, making them "unsatisfactory" job performers in her eyes, some performed very well on the PPT, topping the passing score of 935 by hundreds of points. *See,* PX 61 at EMSCOMB data, lines 77 (low standardized raw criterion/work-sample score; PPT score 1217.67) and line 58 (low standardized raw criterion/work-sample score; PPT score 1164).

116. It is also a fundamental error to define any level of job performance by percentiles or units of standard deviation. Using percentiles or units of standard deviation does not and cannot establish the evidence needed to demonstrate that a passing score represents or predicts

minimally acceptable job performance because minimum acceptable levels of performance depend on absolute not relative performance.[12] They can never validly be defined purely by reference to percentiles or units of standard deviation. As Dr. Gebhardt admitted, the Chicago incumbents in the validation sample were superbly fit. It was irrational to label 15.3% of them as unsatisfactory job performers. Tr. at 669:24-670:5. The only way to figure out how many unsatisfactory workers there are in a workforce is to measure and count—and a plethora of tables, charts, and numbers bound together in a "validation" study cannot change that. It makes no logical or statistical sense to declare, by *ipse dixit*, that a certain number of incumbents *must* be unsatisfactory performers. But that is what Dr. Gebhardt did.

117. As her third and final step toward establishing a passing score for the PPT, Dr. Gebhardt examined "the effect of several cutoff scores on the validity of the acceptances and rejections of incumbents"—with a focus on how many incumbents, despite scoring below their arbitrary cutoff (i.e., in the bottom 15.3% on the work sample tests) would be deemed to pass the PPT at a given cutoff score. DX 37 at ERN000661. On this basis, Dr. Gebhardt recommended implementing a cutoff score of either 984 or 1015. PX 121 at ERN0004841-42.[13] With a passing score set at 984, some 23% of the women in the Chicago incumbent sample would have been deemed to have flunked the PPT. *Id*. With a passing score set at 1015, some 31% of the women in the Chicago incumbent would have been eliminated. *Id.* The City rejected Dr. Gebhardt's recommendations. Probably for reputational reasons, the Fire Department and Chicago legal staff overruled Dr. Gebhardt's advice to use a passing score of either 984 or 1015 and elected to set

[12] *E.g. Evans v. City of Evanston*, 881 F.2d 382, 384-85 (7th Cir. 1989).

[13] ERN0004841 identifies 984 as the recommended passing score; ERN0004842 uses 1015.

the cutoff score, instead, at 935—because, at that score, no Chicago incumbents would be deemed to have flunked. DX 37 at ERN000662.

118. In her description of the method by which the cutoff score was set (DX 37), Dr. Gebhardt does not mention the critical role that both the Fire Department and the legal staff had in determining the cutoff. *See* DX 34 at ERN000661-662. This failure to disclose so material a fact about the rationale and method for setting the cutoff was in violation of the Uniform Guidelines. 29 C.F.R. § 1607.15B(10).

119. There is no evidence that the 935 cutoff was chosen for psychometric reasons. There is no evidence that it was chosen with reference to any actual level of minimum acceptable job performance.

120. Any demonstration of validity based on statistical evidence (so-called "criterion" validity), such as the City attempted in this case, requires "empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance." 29 C.F.R. § 1607.5. B. The City presented no such evidence in this case, either with respect to the "leg lift," "arm endurance," "modified stair climb," or with respect to the 935 cutoff score.

121. The evidence is also very strong that the cutoff was set too high. That is underscored by the fact that at the 935 cutoff, ten percent of the validation sample, 14 New York incumbents, including a startling *one–third* of all female incumbents in the validation sample, all of them employed as big city paramedics, failed the PPT. DX 37 at ERN000657, ERN000658, ERN000662. Absent explanation, it would be irrational to assume that one-third of New York City female paramedics perform their jobs at an unacceptable or unsatisfactory level. That is what Dr. Gebhardt's data and the 935 cutoff imply. Dr. Gebhardt and the city provided no

explanation for these facts, which constitutes a significant failure to prove the job-relatedness of the 935 cutoff score. A cutoff score that flunks one-third of all women in the validation sample, all of them employed as big city paramedics, with no evidence that they are in fact unsatisfactory performers, and no explanation offered for this result, manifestly fails to satisfy the requirement of job relatedness required by Title VII. This creates another hole in the City's evidence which, like others, requires a finding that the City has failed to meet its burden of proof on the issue of job relatedness.

122. The City, although it bears the burden of proof as to the job-relatedness of its cutoff score, offered no testimony by any City official regarding the City's rationale for selecting a 935 cutoff. And there was no testimony that could have helped because there simply is no competent evidence either (a) associating scores below 935 on the PPT with unsatisfactory job performers or (b) associating scores at or above 935 with better job performance. There is also no competent evidence equating a score of 935 with a fair approximation of minimal competence (or any level or paramedic job performance)—all of which follows, inexorably, from the fact that Dr. Gebhardt and HPS never measured actual job performance. Tr. at 668:13-19, 669:5-670:8. There is no evidence justifying the use of 935 as a cutoff score as job related or consistent with business necessity. *Id.*

123. In her final report, Dr. Gebhardt tries to give a veneer of science and objectivity to the specious process by which the cut score was selected. She creates tables and charts and assigns numbers to what were actually arbitrary and statistically nonsensical judgments. She calls this exercise a "utility" analysis. DX 37 at ERN000660-61. It is sophistry and junk science.

124. In the first place, the premise behind this so-called utility analysis is Dr. Gebhardt's arbitrary definition of "unsatisfactory" job performance, which assumed, as

explained above, that 15.3% of incumbents (using a 1.21 standard deviation level to define unsatisfactory performance) must be ineluctably designated unsatisfactory regardless of their actual job performance. One cannot know the rate at which any group of employees performs unsatisfactorily without actually measuring it.

125. Second, Dr. Gebhardt's utility analysis also assumes that PPT scores are linear: that if the PPT were a valid predictor of job performance, which it is not, then higher PPT scores would always predict higher levels of expected job performance. That assertion of linearity is both unproven and doubtful. The City has introduced no competent evidence of a linear relationship between PPT scores and expected job performance. It remains equally or more plausible that, *if* PPT scores predicted paramedic job performance, which they do not, many differences in scores would likely be statistically indistinguishable in terms of the level of job performance they might predict. *See Lewis v City of Chicago,* No. Civ. A1:02 cv 2083, 2005 WL 639618 (N.D. Ill. March 21, 2005), at **4-5. It is also likely that higher scores would be better only up to a point, after which improvements would be redundant, producing only undiscernible changes, if any, in job performance. The City's proof ignores these issues.

126. Third, Dr. Gebhardt's reasoning predicts that based on their scores on the PPT, the plaintiffs in this case are overwhelmingly likely to be "unsatisfactory" paramedics, lacking the ability to perform the essential tasks of the job. DX 37U. Plaintiff Michelle Lahalih, who failed the PPT, was hired by Philadelphia, where she is performing the job. Tr. at 786 8:14-823:8-13. Every one of the plaintiffs in this case testified, without contradiction or any contrary evidence offered by the City, to their ability to perform the essential tasks of the paramedic position. All of the plaintiffs were employed as working paramedics for private ambulance companies or in other jurisdictions when they took the PPT. Dr. Gebhardt's utility analysis

implies that at the time they took the PPT, because they did not score 935 or above, the plaintiffs were not qualified to perform in these paramedic positions they already held, and in many cases had successfully held for years, and that their employers should have terminated them for "unsatisfactory" performance. Unless one presumed that the employers they worked for were keeping them on staff despite their incompetence, an unwarrantedly cynical conclusion for which there is no evidence, the most logical conclusion is that the plaintiffs had the skills and abilities to perform successfully as paramedics and were wrongly screened out by the PPT.[14]

127. Fourth, while a utility analysis purports to measure the marginal utility of a particular cutoff score, it does not provide any answer to the question of what "cutoff score measures the minimum qualifications necessary for successful performance of the job in question." *Lanning v. SE Penn. Transp. Auth.,* 181 F.3d 478, 489 (3d. Cir. 1999).

128. In addition to the psychometrically unjustifiable cutoff score, the City's scoring of the PPT also had other fatal flaws. Dr. Gebhardt had expected that applicants' scores on the "leg lift" test would remain steady or decline slightly with each of their successive trials: "In static strength testing, it is expected to see steady or declining scores on the 2nd and 3rd trial." DX 39 at pp. 7-8. This is true but only under proper test conditions, not the conditions devised by Dr. Gebhardt. To use the machine as a predictive device, it must be pretested in order to determine how many attempts candidates should be allowed to achieve their true performance – which as Dr. McArdle explained should have been evident from the start. PX 23 at pp. 8, 10; Tr. at 441:18-443:25. However, Dr. Gebhardt conducted no pretesting and instead arbitrarily selected three attempts. Candidates were not permitted to practice on these machines before the test.

[14] E.g., *United States v. State of Delaware*, No. CIV.A. 01-020-KAJ, 2004 WL 609331, at *17 (D. Del. Mar. 22, 2004).

Under these circumstances, accurate measurement could not be obtained, particularly for women. This is fatal to any effort to use the machines as a predictive device. *Id.* Gebhardt discovered it only in 2005, after plaintiffs, and many others, had taken the PPT. In 2005, she therefore modified the testing protocol to allow applicants a fourth trial on the "leg lift" in some circumstances. DX 39 at pp. 7-8. The effect of that change in the scoring meant that performance on that event that might lead to a failing score below 2005 could lead to a passing score on the PPT after 2005. Tr. at 1706:13-1707:5.

V. LESSER DISCRIMINATORY ALTERNATIVES

129. The Uniform Guidelines requires employers to give consideration both to: modifying a selection procedure to reduce its disparate impact and (b) replacing a procedure that has disparate impact with another that would have less disparate impact, stating that:

> [W]henever a validity study is called for by these guidelines, the user should include, as a part of the validity study, an investigation of suitable alternative selection procedures *and suitable alternative methods of using the selection procedure* which have as little adverse impact as possible, to determine the appropriateness of using or validating them in accord with these guidelines.

29 C.F.R. § 1607.3.B (emphasis added).

130. In this case, there existed several valid alternative means of using the PPT, which would not have required abandoning it, and also valid alternatives to the PPT, all of which would have resulted in less disparate impact against women. These options were both known to and available to the City. However, the City refused to use them.

131. From the very outset of their relationship, Dr. Gebhardt advised the City that the PPT would have adverse impact against women. PX 54 at ERN002894; Tr. 1668:24-1669:5, 1674:23-1675:3. She also identified available means of reducing that adverse impact. Tr.

1676:18-22. The City took the dangerous course here of hiring Dr. Gebhardt but then dispensing with her advice about already validated means of reducing the disparate impact of the PPT.

132. Because of the near universal passing rate for men (between 2000 and 2009, 98-plus percent of all men who took the test passed it), there is no plausible concern in this case that varying the method of selecting new paramedics might improve the selection rate for both men and women without changing the ratio between them—for example, that providing pre-test physical training to all candidates would benefit all candidates, both male and female, without reducing adverse impact. Simply put, with the male pass rate already at 98-plus percent, there is no real room for it to improve. *See* Tr. at 476:18-19. It is a virtual mathematical certainty that any change that increased applicants' success on the test would accomplish that by improving the female, not the male, pass rate.

A. The City's Refusal to Develop and Implement a Candidate Physical Fitness Program.

133. Dr. Gebhardt specifically proposed that one "method to reduce potential adverse impact is to develop and implement a candidate physical fitness program. Such a program will allow candidates an opportunity to develop the physical abilities needed for acceptable job performance before being test for selection." DX 11 at ERN004399. She noted that HPS "has developed remedial fitness programs" for other clients. *Id.* And she affirmed that these pre-test physical fitness programs work: "It should be noted that these tests *will* have an impact on women." *Id.* (emphasis added).

134. The availability, validity and efficacy of pre-test physical fitness programs toward reducing disparate impact against women was acknowledged at trial by every witness who testified on the subject. Their availability, validity and efficacy is undisputed.

135. Dr. Gebhardt testified that she had created and implemented a physical fitness program in Los Angeles for candidates who were preparing to take a physical performance test for paramedics and that that program worked. Tr. 1679:9-23. She also agreed that research has shown the efficacy of these programs, acknowledging that "a candidate physical training program that is given to candidates before they take a test *has been shown to improve* their likelihood of passing the test." Tr. 1680:19-16 (emphasis added). Indeed, whenever she has instituted such a program she has seen improvement among the candidates who stick with the program, with some candidates who would not have passed but for the training program succeeding as a result of the program. Tr. 1681:17-1682:15.

136. Dr. McArdle agreed with Dr. Gebhardt. He testified that a proper training program works and can improve participants' performance "between 30 and 50 percent, which would definitely help people who are at the lower end of the scoring spectrum;" that if the City had offered a physical training program here, women would have fared better on this test; that the exercise physiology research literature documents that people who have participated in a well-run training program will improve; and, that women are "most definitely help[ed]" by training for physical performance tests. Tr. at 468:4-6, 473:12-14, 474:5-11, 476:20-22; PX 23 (p. 15-16) ("Chicago should have adopted pretest training. Had it done so, this would have reduced or eliminated the adverse impact of the PAT on female applicants . . . [N]o basis exists for [the] speculation that pretest training would not affect the failure relative to male counterparts."). Dr. McArdle also emphasized that PPT had been mischaracterized by Dr. Gebhardt as an "abilities" test. In fact, it was not an "abilities" test but rather a "fitness" test, with one of the differences between fitness and abilities being the comparative ease of improving fitness (but not necessarily ability). Fitness is very amenable to improvement through practice

and training, making it all the more likely that the physical fitness program that Dr. Gebhardt proposed to the City, but which the City refused to implement, would have improved women's success on the test and, thereby, reduced the adverse impact of the test (given the already extraordinarily high pass rate of men). PX 23 (p.7); Tr. 452:23-453:16.

137. Dr. Campion also agreed that pre-test training would have reduced disparate impact. Tr. 673:19-674:9; PX 21 at p. 46.

138. With full knowledge that the PPT would have a disparate impact, the City refused to implement Dr. Gebhardt's recommendation of a candidate physical fitness program as a means of reducing that disparate impact. It is noteworthy that the City has provided no explanation why it refused to offer or coordinate a candidate physical fitness program.

139. The record contains unrebutted evidence that such programs can be implemented without incurring significant expense. PX 23 at pp. 15-16. The feasibility, viability and validity of this option are also illustrated by Los Angeles's having implemented precisely a pre-test training program, created for the city by Dr. Gebhardt. Tr. at 1679:2-23.

140. A candidate physical training program was an available and validated means of reducing the disparate impact of the PPT—if the City had not dispensed with Dr. Gebhardt's advice and refused to use it.

B. The City's Refusal to Provide Applicants a Video With Instructions On How to Train for the Test.

141. As an alternative to providing an actual physical fitness program for candidates, Dr. Gebhardt proposed that the City could reduce the expected adverse impact of the PPT by providing a video to candidates "describ[ing] a fitness program" that candidates "can use to prepare for the test." DX 11 at ERN004399; Tr. 1678:7-21. The City declined to adopt this recommendation. Tr. at 1683:8-10.

142. This alternative was demonstrably available and feasible: the City had provided precisely such a video to candidates taking the City's 1995 Firefighter Test to help them prepare for it. DX 11 at ERN004399. This option would have been a next best alternative to providing an actual training program in fact, and would have aided women and thereby reduced adverse impact for many of the same reasons that an actual training program (or a brochure, as described below) would have been effective.

C. The City's Refusal to Create a Meaningful Brochure, Provided to Applicants Sufficiently in Advance.

143. Dr. Gebhardt also recommended that the City could reduce the expected adverse impact of the PPT by providing a brochure to candidates to help them prepare for the test. Tr. 1683:11-25, 1684:11-19. This is only suggestion that the City took Dr. Gebhardt up on. But the brochure that was used (PX 4) was deficient in material respects, reducing its effectiveness and, in fact, very likely making it counter-productive.

144. First, the research establishes that a brochure should go out at least twelve weeks ahead of test dates, in order to allow candidates a minimum of two to three months to train for the test, which is typically what is required for meaningful strength and aerobic gains. Tr. at 464:9-12. Dr. Gebhardt agreed that a brochure should be mailed at least eight weeks in advance. Tr. at 1683:11-25; PX 23 at p. 11. Here, the City mailed a brochure slightly more than four weeks before testing, which was clearly insufficient and deprived candidates' of information about the test and how to train for it until it was too late. Tr. 1683:11-25, 1684:11-19; PX 23 at p. 11. *See also* PX 4; Tr. 208:6-25, 326:4-14, 809:1-18, 1022:2-10, 1024:9-12.

145. Second, the brochure the City provided to candidates contained significant misinformation. As Dr. McArdle testified, without contradiction (or rebuttal) by any other witness, following the instructions for training provided in that brochure would have misled

candidates. Tr. at 440:21-441:17, 446:6-15. *See also* PX 23 at pp. 8-9, 12. For example, the brochure incorrectly advised that the modified stair climb test was a measure of "muscular endurance." PX 4 at p. 5. In fact it is not. It is a test of cardiovascular fitness, requiring entirely different training than a test of muscular endurance, making the recommendation in the brochure to train for this test by doing exercises to improve the "muscular endurance" in your hands and legs misguided. PX 23 at p. 9. Similarly, the brochure's advice for training for the "arm endurance" test was also misguided. The brochure advised that the ergometer tests "the muscular endurance in your arms," and that training for this test should involve "any activity requiring you to use your arm muscles for an extended period." PX 4 at p. 6. That is simply not the case. Two minutes of all-out pedaling on the ergometer is largely a measure of anaerobic power output, the proper training for which would be repeated *short-term* intervals of intense physical effort of up to 90 seconds duration, not continuous effort for an extended period as recommended by the brochure. PX 23 at p. 9.

146. The undisputed testimony at trial was that a better, more informative brochure, provided eight-to twelve weeks before testing dates, giving candidates *accurate* information about how to train, and sufficient time to train, would have reduced adverse impact against women. PX 23 at p. 11; Tr. 464:17-465:6, 1683:16-25.

D. <u>Additional Available, Appropriate and Lesser Discriminatory Alternatives.</u>

147. In addition to the lesser discriminatory alternatives that Dr. Gebhardt specifically brought to the City's attention, there also existed at least three more obvious and lesser discriminatory alternatives to the PPT.

148. First, candidates would have benefitted from a greater opportunity to familiarize themselves with and use the machines before test day. Physical training is highly specific. Effective training would have required training the specific muscle and metabolic systems used

on the PPT test battery. General gym-type training is not as effective for the highly specific performance demands of the PPT. PX 23 at p. 13-14; Tr. 468:7-469:4.

149. The ideal would have been to set up a test site, with instruction available, for applicants to perform several attempts on the actual test to experience the physical requirements of each test event. *Id.* If Chicago had offered this kind of pretest training opportunity, it would have reduced or eliminated the adverse impact of the PPT on female applicants. PX 23 at p. 15.

150. Second, the City could and should have disclosed to applicants how the test would be scored, and what a passing score would be, before the test, while they were preparing and training for the test. Hiding the passing score from the applicants prejudiced their preparation (PX 23 at pp. 17-18) and, with 98% of all men passing, it is a mathematical certainty that the consequences of prejudicing applicants' preparation prejudiced women more than men, thereby increasing disparate impact. PX 21 at pp. 47-48; Tr. at 471:20-472:9. "It is unacceptable from a professional perspective to train individuals for the physical requirements of a job without knowing what level of acceptable performance is required by the job." PX 23 at pp. 17-18.

151. Third, at the point at which the severe disparate impact of the PPT became clear, and by the time of the results of the second administration of the test were reported in June of 2001, at a minimum, the City could and should have jettisoned the test and returned, at least until a better procedure was found, to the hiring process it had used so successfully for so many years up until the introduction of the PPT. The feasibility of that alternative procedure cannot be questioned: it is proved by the fact that the City had used it, extensively, for years. At the same time, the equal or greater validity of that procedure, compared to the PPT, also cannot legitimately be questioned because there is no competent evidence of the validity of the PPT and no competent evidence of any material weaknesses in the hiring system that the City used for so

many years, so successfully, until April of 2000, when the PPT was first implemented. Throughout the 1980's and 1990's, the City hired large numbers of paramedics through the prior process. As of October of 1999, the CFD was employing 296 paramedics, 172 paramedics in charge, and 59 ambulance commanders, the vast majority of whom had been hired by the CFD through this process during the 1980's and 1990's. DX 37 at ERN000582; PX 14C at ERN002679. They were hired without ever submitting to or passing a pre-employment physical fitness or performance test. And as far as the record discloses, this time-tested method of hiring worked. There is no competent evidence in this record that it did not meet the CFD's legitimate needs. The record, for example, contains no competent evidence—documentary or testimonial, objective or anecdotal—that either public safety or any patient was ever jeopardized by the City's hiring of paramedics, at any point in time, without testing or measuring their physical fitness or performance before hire; that any CFD paramedic hired without taking and passing a physical performance test ever performed unsatisfactorily due to a lack of physical fitness or ability; or that any CFD paramedic has ever been suspended, disciplined, demoted, or terminated due to any inability to perform, or any difficulty in performing, the physical requirements of the paramedic job. And there is *no* evidence that paramedic job performance has improved *on any measure of performance* since the implementation of this test. Tr. at 1454:1-5. There is no evidence that the PPT addressed or cured any problem, only that it created one. A test that creates a barrier to employment for 40 percent of women but only 2 percent of men requires a strong justification under Title VII. There is no such justification in the evidence in this case.

152. There is absolutely no evidence in this record to support the inference that just because the "arm endurance" test involves the arms and the "leg lift" involves leg muscles, and lifting and transporting patients also involves the arms and leg muscles, that success on the "arm

endurance" or "leg lift" test is interchangeable with, or predictive of, or correlated with, the ability to lift or transport patients. That logic would imply that the person who lifts the most weights in the weight room will be the best shot-putter, high jumper, boxer, wrestler or football players. And that is not the case. Each of these tasks requires use of muscles in highly specific and learned movement patterns, which are not dictated or predicted simply by arm or leg strength. This is the principle of "exercise specificity." Tr. at 417:18-21, 418:1-15, 451:14-18. There is no evidence to support the proposition that any component of the final PPT battery, nor the 935 cut-off score, corresponds to a level of physical ability at or below the minimum level required to perform the paramedic job.

153. After years of using the PPT, if that test improved paramedic selection, by any measure, the City should be in a position to demonstrate that. But the City has introduced no evidence of improved paramedic performance as a result of its use of the PPT.

VI. VIOLATIONS OF THE UNIFORM GUIDELINES

154. Although at trial, Dr. Gebhardt and counsel for the City, bragged of her compliance with the Uniform Guidelines, calling them the "gold standard," Dr. Gebhardt's job analysis and validation study in fact failed to comply with at least seven material Guidelines requirements:

a. Section 15.B(2) of the Uniform Guidelines requires that a validation study must contain "an explicit definition of the purpose(s) of the study" and "a description of existing selection procedures." 29 C.F.R. §§ 1607.15.B(2). Dr. Gebhardt's validation study (DX 37) does not discuss the hiring procedures in place before the introduction of the PPT or the purpose of replacing them. *See also* Tr. at 636:15-24 ("[I]f you're going to have an assessment that weeds out half of all women, I'd like to see some evidence that there was a problem that women had

doing the job. And there was no evidence here, so it led to the conclusion that there was not really a solid justification to implement an assessment that had such severe disparate impacts").

b. Section 15.B(5) of the Uniform Guidelines requires that a validation study must contain a "full description of all criteria on which data were collected and means by which they were observed, recorded, evaluated, and quantified," and labels this requirement "essential." 29 C.F.R. § 1607.15.B(5). Dr. Gebhardt's validation study (DX 37) relies on performance by New York City paramedics as a criterion measure, without disclosing who those paramedics are, even that they were from New York, or how their performance was observed, recorded, evaluated and quantified.

c. Section 15.B(6) of the Uniform Guidelines requires that a validation study must provide a "description of how the research sample was identified and selected," labelling that requirement "essential," while also stating that a "discussion of the likely effects on validity of differences between the sample and the relevant labor market or work force is desirable." 29 C.F.R. § 1607.15.B(6). Dr. Gebhardt's validation study (DX 37) does not disclose how either the Chicago incumbents or the New York incumbents in the validation sample were identified and selected. Further, while an early version of Dr. Gebhardt's validation study discussed how the Chicago women in the validation sample "performed better than all or most of the women in other physical performance validation projects," making them atypical (PX 121 at ERN004834), Dr. Gebhardt improperly eliminated that discussion from her final report (DX 37).

d. Section 15.B(8) of the Uniform Guidelines requires that where a criterion-related study, such as Dr. Gebhardt's study in this case, relies on unpublished studies, copies of those studies, or adequate abstracts or summaries, should be attached, labelling that requirement "essential." 29 C.F.R. § 1607.15.B(8). Dr. Gebhardt's validation report relies on several

unpublished technical reports as a basis for claims of validity for several of the components of the PPT final test battery—without attaching any of them.

e. Section 3B of the Uniform Guidelines requires consideration and investigation of both (a) alternatives to the use of a physical performance test, and (b) alternative methods of using the PPT. 29 C.F.R. § 1607.3.B. *See also* 29 C.F.R. § 1607.16.X; Questions and Answers To Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedure, 44 Fed. Reg. 11,996, 12,003 (Mar. 2, 1979), Question No. 48. Dr. Gebhardt has admitted her validation study (DX 37) does not document a search for alternatives. Tr. at 1697:10-12. Dr. Gebhardt and the City violated this requirement of the Uniform Guidelines in this case.

f. Section 14B(4) and 16U of the Uniform Guidelines require that criterion-related validation studies, such as Dr. Gebhardt's, must rely on a representative sample of incumbents (or other subjects), which generates a sufficient range of scores to produce validity results which can be expected to be representative. 29 C.F.R. §§ 1607.14.B(4) and 1607.16.U. *See also* 44 Fed. Reg. 11,996, 12,004 at EEPC Questions and Answers, No. 54. By using non-random, unrepresentative sample of Chicago incumbents, Dr. Gebhardt violated this requirement of the Uniform Guidelines.

g. Sections 14B(2) and 16U of the Uniform Guidelines require the existence of certain key conditions for a criterion-related study, including "having or being able to devise unbiased, reliable and relevant measures of job performance" or other comparably validated, reliable and accurate measures of "work behavior(s) or performance." 29 C.F.R. §§ 9 C.F.R. § 1607.14.B(2) and 16.U. *See also* 44 Fed. Reg. 11,996, 12,004 at EEOC Questions & Answers, No. 54 ("Key conditions for a criterion-related study are a substantial number of individuals for

inclusion in the study, and a considerable range of performance on the selection and criterion measures. In addition, reliable and valid measures of job performance should be available, or capable of being developed. Section 14B(1)"). Dr. Gebhardt used no measures of job performance in this case, and a sample of incumbents who did not exhibit a considerable range of performance, in violation of this requirement of the Uniform Guidelines.

VII. FACTS PARTICULAR TO EACH PLAINTIFF

155. The PPT alone denied each Plaintiff the opportunity to become a CFD paramedic. PX 2A at ERN000295; PX 5A at ERN000364; PX 5B at ERN000319; PX 5C at ERN000340; PX 5D at ERN000388. But for the PPT, Plaintiffs would have been hired by the CFD as paramedics in January 2005. PX 45 at p. 48.

156. When each Plaintiff applied for a CFD paramedic position, the City's 2004 pre-academy requirements for paramedic applicants were a current Illinois EMT/Paramedic license, a current Illinois driver's license, a current Health Care Provider CPR card, and residency in the City of Chicago. PX 1 at ERM000262.

157. Each Plaintiff – Stacy Ernst, Dawn Hoard, Katherine Kean, Michelle Lahalih, and Irene Res-Pullano – met these pre-academy requirements. PX 2A; PX 2B; PX 2C; PX 2D; PX 2E.

158. In addition, each Plaintiff successfully performed the twenty-four tasks that HPS determined were "essential" to CFD paramedics. Tr. at 570:1-572:7.

159. Katherine Kean has been an Illinois-licensed paramedic since 1999. Tr. at 148:11-14. While Ms. Kean was training as a paramedic, she interned for three months at CFD as a paramedic, where she performed well and worked 24-hour shifts. Tr. at 149:14-22, 150:10-16, 155:15-156:13, 158:9-13. Ms. Kean then worked as a paramedic for an ambulance company,

where she was promoted to shift supervisor and worked as a preceptor, reviewing and training newly licensed paramedics. Tr. at 193:1-25.

160. While working for ambulance companies, Ms. Kean responded to a variety of medical emergencies, lifted and extracted people from difficult environments, lifted patients weighing over 245 pounds, and worked during all seasons and storms. Tr. at 159:16-160:22, 163:11-164:6, 166:19-24, 171:11-21. She worked fourteen-hour shifts at least three times per week. Tr. at 150:19-25. In total, Ms. Kean has a total of 14,000 hours of experience. Tr. at 153:8-155:14. Moreover, Ms. Kean taught EMT basics at Malcolm X College in 2002-2003, and taught Wright College and Triton students while working for Medical Express. Tr. at 194:3-21.

161. Dawn Hoard has been an Illinois-licensed paramedic for thirteen years. Tr. at 279:25-280:10. She has significant training. Tr. at 286:5-24. Ms. Hoard's clinical rotations trained her on lifting and maneuvering patients during clinical rotations with Superior Ambulance Company. *Id*. Ms. Hoard's training also included working for the Oak Lawn Fire Department. Tr. at 291:15-292:4.

162. Ms. Hoard has since become a preceptor and instructor at the Christ Hospital Academy, teaching CPR, advanced cardiac life support, prehospital trauma life support, advanced medical life support, and basic lifting and patient extraction. Tr. at 294:1-297:3.

163. Ms. Hoard began working for Dixmoor Fire Department in 2002. Tr. at 301:22-302:8. There, she successfully completed the Fire Academy and became a firefighter paramedic. Tr. at 304:1-12, 304:22-305:16. While working for Dixmoor Fire Department, Ms. Hoard continued learning and training by, for example, attaining certifications in HAZMAT Awareness and HAZMAT Ops. Tr. at 305:20-24. Ms. Hoard responded to a variety of medical emergencies including car accidents, drug overdoses, and gunshot wounds. Tr. at 307:11-308:16. She worked

24-hour shifts, repeatedly carrying patients without ever receiving any complaints or poor performance reviews. Tr. at 309:7-18. Ms. Hoard carried patients as large as 400 pounds. Tr. at 310:6-10. Ms. Hoard worked 16-hour shifts requiring her to lift and carry patients through adverse locations and inclement weather. Tr. at 301:2-21.

164. Michelle Lahalih currently works as a paramedic for the Philadelphia Fire Department. Tr. at 786: 8-18. Previously, Ms. Lahalih worked as a paramedic for an ambulance company, where she also worked 12-hour shifts, transported patients, treated patients, and worked in emergency situations. Tr. at 788:13-17, 791:24-792:12. She worked for the ambulance company for eight to nine years, including during her application to the Chicago Fire Department. Tr. at 795:3-15.

165. While working for the ambulance company, Ms. Lahalih transported patients up and down staircases, including use of a stair chair and a scoop stretcher, and transported patients up to 300 pounds with one partner. Tr. at 798:13-20, 800:1-801:19, 803:15-18. Approximately three-quarters of Ms. Lahalih's work for the ambulance company took place within the city of Chicago. Tr. at 796:4-18.

166. In addition, Ms. Lahalih travelled to New Orleans to treat Hurricane Katrina victims in New Orleans; there, she worked 24-hour shifts. Tr. at 797:1-798:7. She also provided first aid services at Wrigley Field for eleven years, where she had to climb stairs and ramps to treat various injuries. Tr. at 792:20-794:12.

167. Ms. Lahalih also found time to train paramedics. Lahalih was a CPR instructor for six years, training doctors, nurses and first responders. Tr. at 790:12-20.

168. Stacy Ernst is currently the Chief Paramedic for an ambulance company in Chicago. Tr. at 979:20-980:4. There, she not only serves as a paramedic but also as a preceptor

and a field-training officer. *Id*. Ms. Ernst also is involved in the paramedic hiring process for the ambulance company. Tr. at 980:15-25.

169. As an EMT, Ms. Ernst typically worked 10-hour shifts. Tr. at 987:9-16. During a typical 10-hour shift, Ms. Ernst received an average of six or seven calls, all of which required lifting or moving a patient on a stretcher or stair chair. Tr. at 988:1-15. During a typical 24-hour shift with Pulse Ambulance, Ms. Ernst received an average of 14 to 15 emergency and non-emergency calls, including responses to 911 calls. Tr. at 995:17-996:16. She used the same lifting mechanics and techniques regardless of whether a call was emergency or non-emergency. Tr. at 997:15-18.

170. Ms. Ernst's time at paramedic school included clinical time in the emergency room and a minimum 120 hours observation time on an ambulance with a preceptor. Tr. at 991:7-25. Ms. Ernst's preceptor was a female paramedic. Tr. at 993:2-6.

171. As a paramedic, has responded to calls providing medical care while in a moving ambulance. Tr. at 1001:11-17. She has performed CPR, log-rolled patients to stabilize their spine injuries, and used a scoop stretcher to remove patients from difficult environments where a stair chair would not fit. Tr. at 1001:18-1004:21. Ms. Ernst moves patients up and down stairs via a stair chair almost every shift. Tr. at 1004:22-25. She has had to lift and maneuver patients out of cramped spaces. Tr. at 1006:2-11. When responding to calls, Ms. Ernst carried various equipment including a stretcher, stair chair, cardiac monitor and oxygen tank. Tr. at 1011:9-20. Throughout her experience, Ms. Ernst never observed a difference in lifting abilities between the male and female paramedics. Tr. at 993:7-11.

172. Irene Res has worked as a paramedic since 2001 and as an EMT before that. Tr. at 1224:1-3, 1211:18-20. She obtained her EMT-B license at Columbus Hospital. *Id*. at 1207:10-14.

When Ms. Res was training, she received extensive training on paramedic skills, including CPR, lifting and moving patients, and responding to emergencies. Tr. at 1207:17-22. As a paramedic, Ms. Res has had extensive experience working in crowded apartments, homes with broken stairs, and other dilapidated places. Tr. at 1216:3-9.

VIII. PROPOSED CONCLUSIONS OF LAW

173. The claims before the Court arise under the disparate impact provision of Title VII. 42 U.S.C. § 2000e-2(k)(1)(A)(i).

174. Title VII proscribes disparate impact in order to remove "artificial, arbitrary, and unnecessary barriers to employment . . . that had historically been encountered by women." *Connecticut v. Teal*, 457 U.S. 440, 447 (1982) (citing *Griggs v. Duke Power Co*., 401 U.S. 424, 431 (1971)).

175. The litigation and adjudication of Title VII disparate impact claims follows a familiar three-stage analysis. It involves shifting evidentiary burdens, statutorily codified by the Civil Rights Act of 1991. 42 U.S.C. §§ 2000e-2(k)(1)(A)(ii) and (C).

176. At the first stage of this three-stage analysis, "[a] plaintiff establishes a prima facie violation by showing that an employer uses 'a particular employment practice that causes a disparate impact on the basis of race, color, religion, sex or national origin.'" *Ricci v. DeStefano*, 557 U.S. 557, 578 (2009) (quoting 42 U.S.C. § 2000e-2(k)(1)(A)(i)); 42 U.S.C. § 2000e-2(k)(1)(B)(i). To establish this requisite disparate impact, "a plaintiff need only show that [a challenged practice] select[s] applicants for hire in a significantly discriminatory pattern." *Dothard v. Rawlinson*, 433 U.S. 321, 329 (1977); *Thomas v City of Evanston*, 610 F. Supp. 422, 427 (N.D. Ill. 1985). The existence of a "significantly discriminatory" pattern, in turn, may be established either by reference to the EEOC's "four-fifths" test or by tests of statistical significance. *See* C.F.R. § 1607.4(D); *Morgan v. Harris Trust & Sav. Bank of Chi.*, 867 F.2d

1023, 1027 (7th Cir. 1989) (citations omitted). The employer may "point[ ] out deficiencies in [a plaintiff's statistical] data or fallacies in the analysis." *Gulino v. N.Y. State Educ. Dept.*, 460 F.3d 361, 382 (2nd Cir. 2006). However, unless the employer is able to defeat a plaintiff's statistics, the analysis moves to the second stage, at which the burden of proof shifts from the plaintiff to the employer.

177. At the second stage, the employer must rebut and defend against liability "by demonstrating that the [challenged] practice is 'job related for the position in question and consistent with business necessity'"—or else lose. *Ricci*, 557 U.S. at 578; 42 U.S.C. § 2000e-2(k)(1)(A)(i). "Unless and until the defendant pleads and proves a business necessity defense, the plaintiff wins." *Lewis v. City of Chicago, Ill.*, 560 U.S. 205, 213 (2010).

178. In assessing an employer's evidence of job relatedness and business necessity, the touchstone is always whether a challenged practice "bear[s] a demonstrable relationship to successful performance of the job[ ]." *Griggs*, 401 U.S. at 431; 42 U.S.C. §§ 2000e-2(k)(1)(A)(ii) and (C); *Thomas*, 610 F. Supp. at 427. "If an employment practice which operates to exclude [women] cannot be shown to be related to job performance, the practice is prohibited." *Griggs*, 401 U.S. at 431; *Albermarle Paper Co. v. Moody*, 422 U.S. 405 (1975); *Ricci*, 557 U.S. at 578. Unless the employer proves a manifest and demonstrable relationship between the procedure being challenged and successful performance of the job, judgment must be entered for the plaintiff.

179. If, however, the employer carries its burden, then proof and analysis moves to the third stage, with the burden of proof returning to the plaintiff. At this stage, "a plaintiff may still succeed by showing that the employer refuses to adopt an available alternative employment

practice that has less disparate impact and serves the employer's legitimate needs." *Ricci*, 557 U.S. at 578; 42 U.S.C. §§ 2000e-2(k)(1)(A)(ii) and (C).

180. In addition to Title VII, the "Uniform Guidelines on Employee Selection Procedures require reasonable review of valid selection methods which have a disparate impact." *Allen v. City of Chicago*, 351 F.3d 306, 315 n.10 (7th Cir. 2003). The Uniform Guidelines on Employee Selection Procedures are "entitled to great deference." *Albemarle Paper Co.*, 422 U.S. at 431 (quoting *Griggs*, 401 U.S. at 433-34)). The EEOC, Civil Service Commission, and the Departments of Justice, Labor, and Treasury adopted these Uniform Guidelines on Employee Selection Procedures ("Uniform Guidelines").[15] The Uniform Guidelines state that the "use of any selection procedure which has an adverse impact will be considered to be discriminatory and inconsistent with these guidelines, unless the procedure has been validated." 29 C.F.R. § 1607.3(A). If there is an adverse impact, the Uniform Guidelines state that a validity study should include "an investigation of suitable alternative selection procedures and suitable alternative methods of using the selection procedure which have as little adverse impact as possible." 29 C.F.R. § 1607.3(B).

181. "Proof of discriminatory motive . . . is not required under a disparate-impact theory." *Int'l Bhd. of Teamsters v. United States*, 431 U.S. 324, 335 n.15 (1977). "Unlike the disparate treatment theory of liability, which focuses on discriminatory *intent*, disparate impact focuses on discriminatory *consequences*." 1 B. Lindemann, P. Grossman & C.G. Weirich,

[15] *See* 43 Fed. Reg. 38,310 (Aug. 25, 1978) (Civil Service Commission [Office of Personnel Management]); 43 Fed. Reg. 38,311 (Aug. 25, 1978) (Department of Justice); 43 Fed. Reg. 38,312 (Aug. 25, 1978) (EEOC); 43 Fed. Reg. 38,314 (Aug. 25, 1978) (Department of Labor); 43 Fed. Reg. 38,309 (Aug. 25, 1978) (Department of the Treasury). The Uniform Guidelines are codified in 5 C.F.R. § 300.103(c) (2009) (Civil Service Commission [Office of Personnel Management]); 28 C.F.R. § 50.14 (2009) (Department of Justice); 29 C.F.R. § 1607 (2009) (EEOC); 41 C.F.R. § 60-3 (2009) (Department of Labor).

Employment Discrimination Law 3-2 (5th ed. 2012) (footnote omitted); *see Griggs*, 401 U.S. at 432; *E.E.O.C. v. Warshawsky & Co.*, 768 F. Supp. 647, 651 (N.D. Ill. 1991).

182. It is well-established that the Title VII disparate impact standard applies claims brought by an individual plaintiff. *See e.g., Melendez v. Ill. Bell Tel. Co.*, 79 F.3d 661 (7th Cir. 1996); *Santana v. City and Co. of Denver*, 488 F.3d 860 (10th Cir. 2007); *Stockwell v. City and Co. of San Francisco*, 749 F.3d 1107, 1115 (9th Cir. 2014) ("[g]enerally disparate impact analysis is used in a class action, but it may also form the basis of an individual claim") (quoting *Bacon v. Honda of Am. Mfg., Inc.*, 370 F.3d 565, 576 (6th Cir. 2004)). In addition, it is also clear that a prior finding, including a jury verdict, for the employer on intentional discrimination does not preclude the Judge from finding for a Plaintiff on a claim of disparate impact. *See Melendez*, 79 F.3d 661.

183. In this matter, as discussed in more detail below, the PPT had a substantial adverse impact on women. Moving to step two in the disparate impact analysis, the City has failed to carry its burden that the PPT was job-related and consistent with business necessity for three separate reasons: (i) the City failed to demonstrate that the PPT was validated; (ii) even if the PPT was validated, the City failed to justify the setting of the cutoff score and show that it was job-related and consistent with business necessity; and, (iii) even if the PPT was valid and the passing score was justified, the use of the PPT was unlawful because the City failed to search for (and document) alternative practices or uses of the PPT that would have less or no adverse impact.

184. Even if the City carried its burden to show job-relatedness and business necessity, there were less discriminatory alternatives available at the time the PPT was instituted. Both the City's and Plaintiffs' experts agreed that there were available, alternative measures to reduce

adverse impact, including a training program, a video training program, and a brochure. Moreover, the City's prior hiring practice, which did not include the PPT was both substantially equally valid and had less adverse impact.

185. The Court will now proceed to apply these legal principles in more detail to the facts of this case.

IX. PLAINTIFFS' PRIMA FACIE CASE: DISPARATE IMPACT

186. The undisputed facts plainly show that plaintiffs have made out a prima facie case of disparate impact.

187. Plaintiffs carry their initial burden if they "offer statistical evidence of a kind and degree sufficient to show that the practice in question has caused the exclusion of applicants for jobs or promotions because of their membership in a protected group." *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977, 994 (1988).

188. There are at least two widely recognized measures of disparate impact: (1) the EEOC's 80% or "four-fifths" test and (2) tests of statistical significance.

189. The Uniform Guidelines state that a

> selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.

29 C.F.R. § 1607.4(D). The Seventh Circuit and most other courts have "relied on the EEOC's 'four-fifths' rule as a useful benchmark in analyzing disparate impact testing cases." *Davidson v. Citizens Gas & Coke Util.*, 470 F. Supp. 2d 934, 940-41 (S.D. Ind. 2007) (citing *Allen*, 351 F.3d at 310; *Bew v. City of Chicago*, 252 F.3d 891, 893 (7th Cir. 2001); *Cox v. City of Chicago*, 868 F.2d 217, 220 (7th Cir. 1989)).

190. Here, under the 4/5ths rule, where the selection (or pass) rate of women is less than 4/5ths or 80% of the selection rate of men, disparate impact is present. Between 2000 and 2009, the City administered the PPT six times, to some 1,088 applicants for CFD paramedic positions. Over this period, the passing rate for women was 59.18%. By contrast, the passing rate for men was far higher, 98.24%. *See* Proposed Findings of Fact (PFF), *supra*, ¶¶ 4-8. The passing rate for women, was therefore, only 60 % of the passing rate for men, well below the 80-percent standard set by the EEOC to implement the disparate-impact provision of Title VII. 29 C.F.R. § 1607.4.D; *Ricci*, 557 U.S. at 586-87; PX 21 at p 17. In August 2004, when the five plaintiffs in this case took the physical performance test, the disparate impact even greater: the passing rate for women was just 49% of the passing rate for men. PFF, *supra*, ¶ 9.

191. Alternatively, "a *prima facie* case of disparate impact is established by showing through significant statistical disparity that a facially neutral employment practice has a discriminatory impact on a protected class." *Morgan*, 867 F.2d at 1027 (citations omitted).This is often measured in units of standard deviation. The Supreme Court has observed that a difference of "two or three" standard deviations can be sufficient. *Castaneda v. Partida*, 430 U.S. 482, 496 n.17 (1977); *Hazelwood Sch. Dist v. United States*, 433 U.S. 299, 309 n.14 (1977). *Accord: Adams v. Ameritech Servs., Inc.*, 231 F.3d 414, 424 (7th Cir. 2000); *Guardians Ass'n of N.Y. City Police Dep't, Inc. v. Civil Serv. Comm'n of N.Y.*, 630 F.2d 79, 86 (2d Cir. 1980) ("*Guardians I*"). Here, plaintiffs' proof easily surpasses this two-to-three standard deviation standard. Between 2000 and 2009, the pass rates of men and women were separated from the results to be expected from a sex-neutral selection rate by more than 17 units of standard deviation. PFF ¶ 9. In April 2004, the pass rates of men and women were separated by more than 9 units of standard deviation. *Id.*

192. Plaintiffs have demonstrated disparate impact. Further, the magnitude of the disparate impact here is severe: the PPT erected a barrier to employment for 40% of the women applying for the job, but only 2% of men. Women are, on average, twenty times more likely than men to be eliminated in the hiring process by the City's compulsory physical performance test. PPF ¶ 7. This degree of disparate impact calls for heightened scrutiny of the City's assertions of job relatedness. *See Clady v. County of Los Angeles,* 770 F.2d 1421, 1432 (9th Cir. 1985) ("As a general principle, the greater the test's adverse impact the higher the correlation [to job performance] which will be required"); *EEOC v. Atlas Paper Box Co.*, 868 F.2d 1487, 1502 n.24 (6th Cir. 1989) (same); 29 C.F.R. § 1607.14.B(6) ("In determining whether a selection procedure is appropriate for operational use . . . , [t]he degree of adverse impact" should "be taken into account.")

193. Moreover, the PPT's severe adverse impact was readily foreseeable by the City. Prior to the PPT, the test developer told the City that physical abilities tests "*always* result in higher success rates for men than for women . . . ." DX 11 at ERN004393 (emphasis added).

194. Given the PPT's severe adverse impact, as well as its foreseeability, there is enhanced scrutiny of the PPT's validity evidence, the establishment of the cutoff score, and the search (or lack thereof) for alternatives.

X. DEFENDANT'S BURDEN: JOB RELATEDNESS AND BUSINESS NECESSITY

195. In light of plaintiff's proof of disparate impact, the burden shifts to the City to prove that its use of the PPT was job related and consistent with business necessity. 42 U.S.C. § 2000e-2(k)(1)(A)(i). "Unless and until the defendant . . . proves a business necessity defense, the plaintiff wins." *Lewis*, 560 U.S. at 213.

196. An employment test with adverse impact violates Title VII if *either* the abilities measured by the test have not been shown to be related to job performance, *Griggs*, 401 U.S. at 431; *Albermarle Paper Co.*, 422 U.S. at 405; *Ricci*, 557 U.S. at 578, *or* the scoring of the test is not independently job related. It was the City's burden to prove in this case that "all aspects of the test including the method of scoring it" were job related. *Evans v. City of Evanston*, 881 F.2d 382, 384 (7th Cir. 1989). Even if the construction of an exam passes muster, when the exam produces disparate impact, a cutoff scores requires independent and adequate justification. *Id.*; *Guardians I*, 630 F.2d at 106. The City has not carried these burdens in this case. It has failed to prove either the job-relatedness of the three components of the test battery—the "Arm Lift," the "Leg Endurance," and the "Modified Stair Climb"— or the job-relatedness of its cutoff score.

A. <u>The City's Failure To Demonstrate The Job-Relatedness of the Test Battery.</u>

197. Job-relatedness and business necessity, which test developers generally also call "validity," can be demonstrated, under the Uniform Guidelines, by three different kinds of studies: "criterion"-related studies, "content" validity studies, or "construct" validity studies. 29 C.F.R. § 1607.5.A; *Gillespie v. State of Wisconsin*, 771 F.2d 1035, 1040 (7th Cir. 1985).

198. The City made little, if any, attempt to demonstrate the validity of the PPT by arguing that the content of the test—consisting of the "Leg Lift," "Arm Endurance," and "Modified Stair Climb" events—so closely matches and replicates the content of the paramedic job, and is so representative of the range of tasks performed on the job, that its validity can be inferred. In fact, during trial, counsel for the City asked one of plaintiffs' experts to agree that the City had not introduced a content validity study and was not defending its test on that basis. Tr. at 1110:5-13.

199. The City introduced no substantial evidence of "content validity" and failed to establish it.

200. The City also made no attempt to introduce a "construct" validity study, or "construct" evidence, to try validate the PPT on that basis.

201. The City based its job-relatedness defense on an attempt to demonstrate the "criterion" validity of the PPT. By casting its lot with "criterion" evidence, the City undertook, therefore, to demonstrate with empirical evidence that scores on the PPT were "predictive of or significantly correlated with important elements of [paramedic] job performance." 29 C.F.R. 1607.5.(B) ("Evidence of the validity of a test or other selection procedure by a criterion-related validity study should consist of empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance").

202. For the criterion-related study, the City relied on a single criterion, a so-called "work sample" test. DX 37 at ERN000635, ERN000647, ERN000055-62.

203. The City's evidence of job-relatedness came up short for five reasons, each of them independently dispositive of this case, each by itself requiring, without more, entry of judgment in plaintiffs' favor. First, the work sample tests did not measure important elements of work performance. Second, the measures for scoring the work sample tests were unrelated to important elements of work performance. Third, the sample of incumbents used in the development of the criteria was unrepresentative and cannot support validity. Fourth, the criterion focused on a narrow part of the job and overemphasized physical abilities. Fifth, the absence of any measure of the reliability of the PPT.

204. Further, the City has not carried its burden requiring the PPT's cut off score.

205. Lastly, although the City was required to search for less discriminatory alternatives and to document this search, the City never undertook such a search.

1. No Measure of Actual Job Performance.

206. The first hole in the City's evidence, fatal in itself, is the absence of any measure of actual paramedic job performance. It is axiomatic that validating a pre-employment test by means of a criterion-related study requires "having or being able to devise unbiased, reliable and relevant measures of job performance." 29 C.F.R. § 1607.16.U. Accordingly, the "most important" part of a criterion-related study is a "good measure of job performance." Tr. at 655:18-20.

207. Here, however, the City never compared performance on the PPT to a reliable or relevant measure of job performance. PFF ¶ 51. And without such a measure, the City did not and could not successfully validate the PPT or adhere to the Uniform Guidelines. The absence of an accurate, or any, measure of actual job performance presents an insuperable obstacle to establishing the criterion validity of an employment test, constituting a gaping hole in the City's evidence in this case.

208. Rather than any measure of actual job performance, Dr. Gebhardt resolved to use scores of incumbents on several work "sample" tests as a criterion measure instead. This meant using scores on one set of tests (the "work samples") to try to validate another *tests* (the PPT) without ever bothering to measure *actual job performance*. Both conceptually and in its execution this strategy was a failure. When the thing to be predicted (the criterion) is job performance, there is no substitute, or legitimate way around measuring actual job performance. To establish the validity of a pre-employment test by means of criterion evidence, the criterion must be a measure of actual job performance, not scores on another test, which itself has never been validated against *job performance.*

209. The City would have the court find job-relatedness on the basis of an asserted correlation between the results of two separate sets of tests (the work "sample" tests and the three

"abilities" tests included in the final PPT battery), both devised by Dr. Gebhardt. However, *neither* set of tests was ever validated as, or can substitute for, a measure of actual job performance. The Court cannot accept the flawed argument that the job-relatedness of the PPT could be established by correlating PPT scores with scores on another set of physical tests. Correlation to actual job performance remains unaccomplished. By substituting performance on the work "samples" as a correlate to performance on the PPT, the City cannot avoid the obligation to show job relatedness—meaning a correlation between the PPT and actual *job performance.* Since the work sample tests do not replicate requirements of the job and are not actual "important elements of job performance," these tests are not an appropriate criterion. PX 21 at pp. 33-34; PX 22 at pp. 29-30.

210. As the district court wrote in *Guardians,* while emphasizing a similarly fatal hole in the City of New York's proof of job-related in that case:

> Defendants would have this court find job-relatedness on the basis of a high correlation between the results of two separate testing practices, neither of which by itself has been validated according to accepted methods. We cannot accept this flawed argument.

Guardians Ass'n of New York City Police Dep't, Inc. v. Civil Serv. Comm'n of City of N.Y., 431 F. Supp. 526, 547 (S.D.N.Y. 1977), *vacated and remanded on other grounds sub nom.*, 562 F.2d 38 (2d Cir. 1977). In reviewing that ruling the Second Circuit emphatically agreed, stating:

> [The district court correctly reasoned that] "[d]efendants would have this court find job-relatedness on the basis of a high correlation between the results of two separate testing practices, neither of which by itself has been validated according to accepted methods. We cannot accept this flawed argument." ***Nor can we.***

(*Emphasis added*). *Guardians Ass'n of N.Y. City Police Dep't, Inc. v. Civil Serv. Comm'n of City of N.Y.*, 633 F.2d 232, 244 (2d Cir. 1980) ("*Guardians II*"), *aff'd sub nom.*, 463 U.S. 582 (1983).

2. The Measures for Scoring the Work Sample Tests Were Unrelated to Important Elements of Work Performance.

211. The difference between the work sample tests and the tasks actually performed on the job by paramedics is enhanced by the fact that scores on the work sample tests were determined on a basis that was unrelated to the requirements of actual job performance. The score on the "lifting test" was based upon the amount of weight that a member of the incumbent Chicago sample was able to lift over 13 cycles of the test where 10 pounds was added in each cycle up to a maximum of 210 pounds. DX 37N at ERN 005696-98. The score for both the "equipment carry" and "stair push" tests was based upon the time that it took the participants to complete the tests. *Id.* at ERN005702 (equipment carry), ERN005706-07 (stair push).

212. By scoring these work sample tests on the basis of "speed" and total lifting capability, the tests resemble "maximum performance" tests or "athletic contests." Tr. at 664:19-666:25. By basing the scoring on, maximum performance, without any demonstrable need for this maximum performance, this criterion is ill-suited to establish the minimum qualifications that is the basis for a cutoff score on an entry level test.

213. In describing the physical tasks and physical abilities in her Report, Dr. Gebhardt did *not* provide any foundation for evaluating or scoring the work sample tests on the basis of "speed" or relative lifting capacity. DX 37 at ERN000588-93. Indeed, there is no reference to "speed" on the paramedic task list nor is there any reference to lifting 210 pounds, the maximum weight on the lifting test. *Id.* at Appendix C. Moreover, Dr. Gebhardt had previously "concluded that qualitatively, a single person lift of 150 pounds of the [pertinent] apparatus should be deemed a passing score for a paramedic applicant." PX 21 at p. 36 n.21.

214. There was no evidence in the job analysis regarding the speed by which a paramedic needs to go up and down stairs. Tr. at 666:15-17; Tr. at 1752. Similarly, in the job

analysis, Dr. Gebhardt did not ask about the speed by which a paramedic pushes a stair chair. *Id.* This is an especially glaring omission because Dr. Gebhardt did not evaluate the speed of paramedics performing the job in Chicago although she had done so in other studies. Tr. at 1747:21-1748:8.

3. The Absence of a Representative Sample of Incumbents, Which, as a Matter of Law, Renders Dr. Gebhardt's Statistics Unreliable.

215. The second fatal flaw in the City's attempted demonstration of job-relatedness is also incurable and arises from the flaws in the incumbent sample that Dr. Gebhardt used for purposes of her validation study. It is axiomatic that an analysis is only as good as the data on which it rests and that "[i]nferences from the part to the whole are justified only when the sample is representative." Reference Guide on Statistics, in Reference Manual on Sci. Evid. 211, 217 (Fed. Jud. Center) (3d ed. 2011). These conditions were violated here. PFF ¶¶ 70-81.

216. The formation and composition of Dr. Gebhardt's sample population, whose scores on her work "sample" tests and on the PPT were supposed to stand in for a criterion measure of actual job performance, did not comport with any scientific sampling standard. The group was not randomly selected. It was a "convenience" sample. *DeKoven v. Plaza Assocs.*, 599 F.3d 578, 581 (7th Cir. 2010) ("convenient to the sampler—as distinct from a 'representative' sample—representative of the population sampled."). The City selected a group of paramedics and then invited volunteers, who were neither comparable to nor representative of the larger overall population of CFD paramedics. PFF ¶¶ 73-81.

217. All of the statistical tables and charts from which the City asks the Court to infer the validity of the PPT, based on the test scores of incumbents, are drawn from this non-random, unrepresentative sample of incumbents. The fact that this "validation" group was neither randomly drawn from the population of CFD incumbent paramedics nor representative of the

population of CFD paramedics (or the labor pool) introduced incurable error into the attempt to validate the PPT. As a matter of law, it renders Dr. Gebhardt's use and reliance on data drawn from this sample statistically unreliable and inconsistent with accepted statistical methods and principles and Fed. R. Evid. 702. *Camesi v. Univ. of Pittsburgh Med. Ctr.*, Civil Action No. 09-85J, 2011 WL 6372873, at *11 (W.D. Pa. Dec. 20, 2011) ("[T]he absence of random sampling renders [the expert's] opinions unreliable"). *See also United Parcel Serv., Inc. v. U.S. Postal Serv.*, 184 F.3d 827, 840 (D.C. Cir. 1999) ("[V]alidity . . . is undermined if the sample is not representative of the population it purports to represent or is not selected in a sufficiently random manner."); *In re Chevron U.S.A., Inc.*, 109 F.3d 1016, 1019-20 (5th Cir. 1997) ("The sample must be a randomly selected one of sufficient size . . . Such samples are selected by the application of the science of inferential statistics. The essence of the science of inferential statistics is that one may confidently draw inferences about the whole from a representative sample of the whole . . . . It is [only the assurance that the sample was representative] that [can] provide the foundation for any inferences that may be drawn . . . . Without [it] . . . , no inferences may be drawn").

218. Any standard-setting exercise, including to validate a test or set a cutoff score for a test, where the basis for setting the standard involves the performance of only a sample of the population of interest, requires a representative sample of that population. *United States v. City of Erie, Pa.*, 411 F. Supp. 2d 524, 562-67, 570 (W.D. Pa. 2005) ("While this Court does not question the City's good faith . . . , the decisions made by the City in administering that exercise and in choosing the cut-off score for new officer applicants made it probable that the PAT's passing standard would be set at an inappropriately high level, both because the City used a non-representative sample of 19 volunteers and because the City chose to utilize the average scores of

the volunteers, all of whom the City admitted were performing their jobs at least adequately, rather than determining the level which distinguished successful from unsuccessful performers.").

219. Because Dr. Gebhardt's "validation" sample of incumbents was not randomly selected or representative, her results, based on the test performance of those incumbents, both on the PPT and on her work "sample" tests, are statistically unacceptable, unreliable, and inadmissible. PFF ¶¶ 73-81; *United Parcel Serv.*, 184 F.3d at 840; *In re Chevron*, 109 F.3d at 1019-20; *Camesi*, 2011 WL 6372873 * 11; *City of Erie, Pa.*, 411 F. Supp. 2d at 570.

220. Without competent evidence, drawn from a representative sample of incumbents, from which to validate the PPT on the basis of the performance of those incumbents on it, the City has no criterion evidence—either of incumbents' actual job performance, as Title VII requires, *Guardians II*, 633 F.2d at 244, or even of their performance on the work sample tests that Dr. Gebhardt tried to use as a stand-in for a measure of actual job performance. This is a gaping hole in the City's evidence, dispositive of the case, dooming its job-relatedness defense, and requiring the entry of judgment in favor of the Plaintiffs.

221. Because there is no evidence that the incumbents whose test performance forms the basis for Dr. Gebhardt's inferential statistics were appropriately representative of the larger population of CFD paramedics whose level of job performance she was seeking to predict, and overwhelming evidence, in fact, demonstrating that they were not (PFF ¶¶ 73-81) the Court cannot and will not accept Dr. Gebhardt's statistical evidence as competent evidence of the job-relatedness or business necessity of the PPT.

4. The Criterion is Focused on a Narrow Part of the Job and Overemphasizes Physical Abilities.

222. The Uniform Guidelines provide that there is "close review" when there is reliance "upon a single selection instrument which is related to only one of many job duties or aspects of job performances will also be subject to close review." 29 CFR Section 1607.14.B.(6). Furthermore, the "Principles for the Validation and Use of Personal Selection Procedures," Society for Industrial and Organizational Psychology, (4th Ed. 2003) at 16, provides that a "criterion measure is deficient to the extent that it excludes, relevant, systematic variance [such as] work behaviors or outcomes critical to job performance." Accordingly, validation studies "should not focus on only predicting a narrow slice of the job." PX 22 at p. 31.

223. Here, the selection instrument developed by the City focuses only upon some of the physical aspects of the job, a narrow slice of the paramedic job, which includes many other important aspects, such as medical knowledge, relating to and dealing with diverse population and difficult situations, and technical proficiency. PX 21 at p. 25.

224. Prior to 2000, the City did not use any physical abilities test and there was no evidence that paramedics were unable to do their job. Tr. at 636:13-18. *See* also DX 11. There is no evidentiary basis submitted by the City describing any problem with the physical ability of the current paramedics in Chicago to perform the job successfully even though they were hired without any physical ability test. PFF ¶ 5. Thus, it is unsurprising that Dr. Gebhardt – not the City – pushed for a physical abilities test, which she did prior to any job analysis of CFD paramedics. DX 11 at ERN004377-78. For these and other reasons, the job analysis was conducted in a biased manner. Tr. at 633:2-7.

225. It is particularly problematic that the process relied upon by the City led to a focus solely of physical abilities when it is well-known, as the City's expert plainly stated, that "due to

physiological differences between men and women . . ., physical performance testing will *always* result in a higher success rate for men than for women." DX 11 at ERN004393 (emphasis added). Because the PPT eliminated approximately half of the women, it was especially important to have evidence that there was a problem with women incumbent paramedics. Tr. at 636:15-24.

226. The validation study, itself, provides an illustration of the manner by which the City focused solely on physical abilities in an unjustified and biased manner. Following the analysis of the job tasks of paramedics, the City determined that it was appropriate to examine whether "personality attributes" and "physical abilities" were essential or not to the performance of the paramedic job. DX 37 at ERN000579, ERN000594-95. Dr. Gebhardt determined that nine personality attributes were identified as essential by a majority of the raters, that is, those attributes received a score of 2 or "essential" on a three-point rating scale used for these personality attributes. *Id.* at ERN000596. The ratings for the personality attributes were comparable to or greater than the ratings determined for the physical abilities, which received a rating of 3.5 or "moderate" on the seven physical abilities. Yet the City adopted a selection test based only upon the physical abilities, which would have a foreseeable severe adverse impact on women, but not for the personality attributes, which was not shown to have adverse impact on women.

5. <u>The Absence of Any Measure of the Reliability of the PPT, Which Precludes Any Finding of Validity or Job-Relatedness.</u>

227. The reliability of the PPT is unknown. PFF ¶¶ 101-104. The absence of an accurate, or any, measure of the reliability of the PPT is dispositive of the City's attempt to establishing the validity of the PPT: without evidence establishing reliability, a finding of validity is not possible. *Gillespie*, 771 F.2d at 1041 (a criterion-related test "must be tested for

reliability"); *United States v. State of Delaware*, No. 01-cv-020-KAJ, 2004 WL 609331, at *6 (D. Del. Mar. 22, 2004) ("Reliability of a test is a necessary condition for validity") (citation omitted).

B. The City's Failure to Demonstrate the Job-Relatedness of the Cutoff Score.

228. The City's failure to demonstrate the job-relatedness of its cutoff score of 935 constitutes reflects still another fatal hole in the City's job-relatedness defense—and a fourth independently dispositive basis requiring rejection of that defense and the entry of judgment in plaintiffs' favor.

229. It is the employer's burden to prove that "all aspects of the test including the method of scoring it" are job related. *Evans*, 881 F.2d at 284 (7th Cir. 1989). The City has not done that here.

230. "When a cutoff score unrelated to job performance produces disparate . . . results, Title VII is violated." *Guardians I*, 630 F.3d at 105; *Thomas*, 610 F. Supp. at 431.As Dr. Gebhardt acknowledged, professional standards require setting a cutoff score for a pre-employment test require setting a "physical test cutoff score[ ] . . . [at] the minimum acceptable level." DX 37 at ERN000660; Tr. at 1732:1-5. PFF ¶ 108. *See also Lanning v. Se. Pa. Transp. Auth.*, 181 F.3d 478, 481 (3rd Cir. 1999) ("a discriminatory cutoff score on an entry level employment examination must be shown to measure *the minimum* qualifications necessary for successful performance of the job in question in order to survive a disparate impact challenge.") Even if the City had met its burden of showing that the PPT battery was job-related, it still did not and could not meet its burden of showing that the cutoff score was established property.

231. First, the effort to establish a defensible cut off score on the PPT was futile from the outset. Given the invalidity of each of the three tests that make up the PPT, which do not predict and are not significantly correlated with actual job performance, there is no passing score

methodology that could have salvaged the PPT. These facts compel the conclusion that neither the PPT nor the cut scores were job related.

232. Second, establishing a cutoff at a level predictive of successful job performance, much less at the level of the "minimum qualifications" required for successful performance, *Lanning*, 181 F.3d at 481, where Dr. Gebhardt acknowledged it should be set, presumes the existence of a measure of job performance. A cutoff score cannot be correlated to any particular level of job performance when job performance has not been measured. As discussed above, the City never measured actual job performance. PFF ¶ 108.

233. Third, the City also doomed any attempt to defend its cutoff score as job-related when it elected to establish that score by reference to percentiles or units of standard deviation. PFF ¶ 116. Cutoff scores can never be validly established purely by reference to percentiles or units of standard deviation. The cutoff score here, set so that 15.3% of incumbents would "fail," is no more valid than the cutoff score that was rejected by the court in *Thomas v. City of Evanston*, which was set so that 16% of incumbent officers would "fail." 610 F. Supp. at 431. All cutoff scores based on percentiles or units of standard deviation are invalid, as a matter of law, for the reasons that Judge Posner explained in *Evans v. City of Evanston*. Choosing a passing score one standard deviation above—or below— the mean is never job-related. 881 F.2d at 384. It cannot be job related—because, as Judge Posner explained, "the ability to perform firefighting [or paramedic] tasks adequately depends not on relative but on absolute test performance," whereas a cutoff score set with reference to percentiles or units of standard deviation *always* refers to relative performance. *Id.* Further, exactly the danger that Judge Posner identified—"if one year all the applicants were superbly fit, it would be irrational to disqualify the entire bottom 16 percent"— materialized in this case. *Id*. Dr. Gebhardt's "validation" sample of Chicago

incumbents was extraordinarily physically fit. PFF ¶¶ 75-78. For the reasons that Judge Posner explained, it was irrational and not job-related to set a pass score that would disqualify the entire bottom 15.3% of that sample—and 40% of all women applicants. *See also, Thomas*, 610 F. Supp. at 431 ("Because the City asserts that a person who flunks the test would be an incompetent police officer, one would expect evidence that at least 16% of incumbents who 'flunk' are not in fact performing their jobs satisfactorily. The City has presented no evidence to that effect.").

234. After *Evans v. City of Evanston* and *Thomas v. City of Evanston*, the fallacy of setting a cutoff using percentiles or units of standard deviation is beyond legitimate dispute. Such cutoff scores are not job related. *Thomas*, 610 F. Supp. at 431. ("[The employer] established its norms by giving the test to incumbent police officers. It scaled the test so that 16% of incumbent officers would fail. The score on each subtest was scaled so that any person scoring within one standard deviation of the mean would receive a passing score of '3' for that test . . . . The City has simply not presented evidence that this scaling system validly predicts future performance . . . [N]o evidence supports what appears to be pure speculation that . . . an applicant who scores in the lowest 16% of incumbent officers would perform incompetently").

235. Fourth, any scoring system that functions as Dr. Gebhardt's system for scoring the PPT did, by summing weighted scores from multiple sub-tests, "allows for an individual to compensate for a low score on one test with a higher score on another test." DX 37 at ERN000657. By definition, that makes it possible to pass without "attain[ing] a[ny] specific score on" any one of the tests in the test battery. *Id.* As a result, this "compensatory" model for scoring necessarily implies: (a) that none of the abilities being tested for are essential to job performance, since otherwise it would be unacceptable to score the test in a manner that allows a

superior performance on one to compensate for poor performance on another; and, (b) if they are essential, then the passing score does not represent or predict a minimally acceptable level of job performance because it does not insure any minimum level of ability on any of the measurements being taken. These facts also independently compel the conclusion that PPT's cutoff score of 935 was not job related.

C. The City Failed to Search (and document this search) for Less Discriminatory Alternatives When the PPT was Developed.

236. As part of performing a validity study, a user should examine alternative selection procedures or uses for the procedure that has less or no adverse impact. If there are two substantially equally valid procedures or uses, the user should employ the procedure or use that has less adverse impact. This search should be documented. Chicago undertook no such search. The developer of its selection procedure did not document any such search.

237. The Uniform Guidelines state that:

> W]henever a validity study is called for by these guidelines, the user should include, as a part of the validity study, an investigation of suitable alternative selection procedures and suitable alternative methods of using the selection procedure which have as little adverse impact as possible, to determine the appropriateness of using or validating them in accord with these guidelines."

29 C.F.R. § 1607.3.B. Under the Guidelines, "a validity study is called for" whenever disparate impact exists. Under the above-quoted section of the Guidelines, the existence of disparate impact, present here, imposed two sets of duties on the City: (a) a duty to investigate and consider "alternative selection procedures," i.e., by replacing the PPT; and, (b) "alternative methods of using the [same] selection procedure," meaning, in this case, continuing to use the PPT but after adopting alternative means of preparing applicants for it or of scoring it. *See also* 29 C.F.R. § 1607.16X; Questions and Answers To Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedure, 44 Fed. Reg. 11,996, 12003 (Mar.

2, 1979), Question No. 48 ("Q. Do the Guidelines call for a user to consider and investigate alternative selection procedures when conducting a validity study? A: Yes . . .").

238. Moreover, two Circuits have clearly underscored this obligation to search for suitable alternative selection procedures that would reduce or eliminate significant adverse impact. *Officers for Justice v. Civil Service Comm'n*, 979 F.2d 721, 728 (9th Cir. 1992) ("before utilizing a procedure that has an adverse impact on minorities, the city has an obligation pursuant to the *Uniform Guidelines* to explore alternative procedures and to implement them if they have less adverse impact and are substantially equally valid to [the proposed procedures]," *citing* 29 C.F.R. § 1607.3B.); *Brunet v. City of Columbus*, 1 F.3d 390, 412 (6th Cir. 1993), *cert. denied*, 510 U.S. 1164 (1994) ("before implementing a process . . . that has an adverse impact on women, the City is obligated to *conduct its own investigation of viable alternatives* with lesser or no impact on female applicants . . . . There is nothing in the record indicating the city *explored alternatives* to strict rank ordering or that the District Court looked to see that the City had done so; We believe this was error."); *Brunet v. City of Columbus*, 58 F.3d 251, 254, 259 (6th Cir. 1995) (same).

239. Under § 1607.3.B of the Uniform Guidelines and the caselaw, quoted above, even if the PPT had been job-related, and even if the cutoff score had been properly validated, the City's use of the PPT would still have been unlawful, given the manifest disparate impact it caused, because, as the evidence shows, the City both: (a) failed to investigate suitable alternatives to the PPT and (b) failed to investigate suitable alternative methods of using the PPT. PFF ¶¶ 129-151.

XI. LESSER DISCRIMINATORY ALTERNATIVES

240. Given the City's manifest failure to demonstrate the job relatedness of the PPT or its cutoff score, it should not be necessary to reach the third prong in a Title VII disparate

analysis—the existence and availability of lesser discriminatory alternatives to the PPT. 42 U.S.C. §§ 2000e-2(k)(1)(A)(ii) and (C). However, it is addressed here for the sake of a complete record.

241. Plaintiffs have amply carried the burden of demonstrating that: (a) there were several valid alternative means of using the PPT, which would not have required abandoning it, and would have resulted in less disparate impact against women: (b) there were also valid alternatives to the PPT, which would have resulted in less disparate impact against women; and, (c) these options were both known to and available to the City, which, however, refused to investigate, consider, use or adopt them. PFF ¶¶ 133, 141-46. By failing to adopt these legitimate alternatives, the City violated Title VII.

242. First, the availability, validity and efficacy of pre-test physical fitness programs toward reducing disparate impact against women was acknowledged at trial by every witness who testified on the subject. Their availability, validity and efficacy is undisputed. PFF ¶¶ 133-40. The City provided no explanation, let alone evidence, as to why it refused to offer or coordinate a candidate physical fitness program. Such a program was an available and validated means of reducing the disparate impact of the PPT that the City was required, by virtue of Title VII, to adopt. Dr. William McArdle, an expert in exercise physiology, testified that a proper training program can improve performance between 30 and 50 percent. Tr. at 468:4-5.

243. Second, the availability, validity and efficacy of providing a video to candidates describing a fitness program they could use to prepare the test has also been demonstrated. Indeed, the City had provided precisely such a video to candidates taking the City's 1995 Firefighter test. PFF ¶¶ 141-42.

244. Third, Plaintiffs have also shown the existence of an obvious and equally valid, less discriminatory alternative to the City's use of the PPT: not using it. Courts may infer feasibility where the employer previously used Plaintiffs' proposed procedure. *See Easley v. Anheuser-Busch, Inc.*, 572 F. Supp. 402, 410 (E.D. Mo. 1983) *aff'd in part, rev'd in part*, 758 F.2d 251 (8th Cir. 1985) ("An alternative selection procedure, without the same adverse impact on black applicants, was available to the defendant," namely, the same hiring procedure that defendant followed before instituting formal testing, which, as far as the evidence showed, resulted in hires who performed their jobs in a satisfactory manner).

245. The feasibility of resuming the hiring procedure that existed for so many years before 2000: it is proved by the fact that the City had used it, extensively, for years. At the same time, the equal or greater validity of that procedure, compared to the PPT, also cannot legitimately be questioned because there is no competent evidence of the validity of the PPT and no competent evidence of any material weaknesses in the hiring system that the City used for so many years, so successfully, until April of 2000, when the PPT was first implemented. Throughout the 1980's and 1990's, the City hired hundreds of paramedics through the prior process. As of October of 1999, the CFD was employing 296 paramedics, 172 paramedics in charge, and 59 ambulance commanders, the vast majority of whom had been hired by the CFD through this process during the 1980's and 1990's. DX 37 at ERN000582; PX 14C at ERN002679. They were hired without ever submitting to or passing a pre-employment physical fitness or performance test. And as far as the record discloses, this time-tested method of hiring worked. There is no competent evidence in this record that it did not meet the CFD's legitimate needs. The record, for example, contains no competent evidence—documentary or testimonial, objective or anecdotal—that either public safety or any patient was ever jeopardized by the

City's hiring of paramedics, at any point in time, without testing or measuring their physical fitness or performance before hire; that any CFD paramedic hired without taking and passing a physical performance test ever performed unsatisfactorily due to a lack of physical fitness or ability; or, that any CFD paramedic has ever been suspended, disciplined, demoted, or terminated due to any inability to perform, or any difficulty in performing, the physical requirements of the paramedic job. And there is *no* evidence that paramedic job performance has improved *on any measure of performance* since the implementation of this test. Tr. at 1454:1-5. There is no evidence that the PPT addressed or cured any problem, only that it created one. A test that creates a barrier to employment for 40 percent of women but only 2 percent of men requires a strong justification under Title VII. There is no such justification in the evidence in this case.

246. Plaintiffs also proved the existence of several other alternative means of utilizing the PPT that would have had less disparate impact, including a video with instructions on how to train and prepare for the test (recommended by Dr. Gebhardt but which the City refused to adopt), see PFF ¶¶ 141-42; a useful test brochure provided at least eight to twelve weeks in advance of the test, rather than the brochure with erroneous training instructions provided only four weeks before the test (*see* PFF ¶¶ 143-46) opportunities to practice on the test machines in advance, which Dr. McArdle testified would have reduced adverse impact (Tr. at 437:7-8, 443:7-25, 453:14-16; PX 23 at p. 10; PFF ¶¶ 137-39); and disclosure of the PPT scoring mechanism and passing score, which Dr. McArdle also testified was required for candidates to effectively prepare and would have reduced disparate impact. Tr. at 471:20-472:9; PX 23 at pp. 17-18. The City failed to refute any one of these alternative means of using the test.

XII. REMEDIES

247. Each Plaintiff is entitled to relief because the City's unlawful discrimination prevented each Plaintiff from being hired as a CFD paramedic. One of "the central purposes of Title VII is to make persons whole for injuries suffered on account of unlawful employment discrimination." *Franks v. Bowman Transp. Co.*, 424 U.S. 747, 763-64 (1976) (internal quotations and citation omitted). Because there is a disparate impact finding, each Plaintiff is entitled to equitable relief. 42 U.S.C. § 2000e-5(g). This equitable relief includes back pay and instatement. *Id*. ("the court may . . . order such affirmative action as may be appropriate, which may include, but is not limited to, reinstatement or hiring of employees, with or without back pay . . . or any other equitable relief as the court deems appropriate."); *see also Franks*, 424 U.S. at 763-64.

248. A Title VII disparate impact finding creates a "strong presumption" that "can seldom be overcome" in favor of back pay. *Thomas v. City of Evanston*, 610 F. Supp. 422 (N.D. Ill. 1985) (quoting *Liberles v. County of Cook*, 709 F.2d 1122, 1136 (7th Cir.1983). In other words, an award of back pay does not require a finding of intentional discrimination. *E.g.*, *Liberles*, 709 F.2d at 1135 (it is "well settled that retroactive relief or back pay awards against state and local governmental entities who unlawfully discriminate, either through disparate impact or disparate treatment, against their employees is appropriate") (collecting cases); *McReynolds v. Merrill Lynch, Pierce, Fenner & Smith, Inc.*, 672 F.3d 482, 483-84 (7th Cir. 2012) (stated that the plaintiffs could obtain equitable relief, including back pay, without proof of intentional discrimination in a Title VII action). For example, in *Thomas*, the Court awarded back pay to female plaintiffs who suffered disparate impact but not intentional discrimination as a result of a physical abilities test. 610 F. Supp. 422.

249. The purpose of back pay under Title VII is "to restore victims of discrimination to a position where they would have been were it not for the unlawful discrimination." *Watson v. Potter*, 2002 WL 31006129, at *5 (N.D. Ill. Sept. 5, 2002) (citing *Franks*, 424 U.S. at 763-64). In order to make whole the victims of discrimination, the back pay award must include lost benefits. *Curler v. City of Fort Wayne*, 591 F. Supp. 327, 338 (N.D. Ind. 1984) (citing *Bowe v. Colgate-Palmolive Co.*, 489 F.2d 896 (7th Cir.1973)); *see also Metz v. Merrill Lynch, Pierce, Fenner & Smith, Inc.*, 39 F.3d 1482, 1493 (10th Cir. 1994); *Wirtz v. Kansas Farm Bureau Servs., Inc.*, 274 F. Supp. 2d 1215, 1218 (D. Kan. 2003). For the same reason, parties "who prevail on their Title VII . . . claims are entitled to prejudgment interest on their back pay award." *Lalowski v. Corinthian Schools, Inc.*, 2013 WL 1788353, at *7 (N.D. Ill. Apr. 26, 2013).

250. Back pay relief is calculated by measuring "the difference between actual earnings for the period and those which the plaintiff would have earned absent the discrimination by defendant." *Geraty v. Vill. of Antioch*, 2014 WL 1475574, at *1 (N.D. Ill. Apr. 15, 2014) (internal quotations omitted) (citing *Horn v. Duke Homes, Div. of Windsor Mobile Homes, Inc.*, 755 F.2d 599, 606 (7th Cir. 1985)). Unsurprisingly, this calculation is seldom made with exact precision. *See Geraty*, 2014 WL 1475574, at *2 ("measuring back pay always involves some level of speculation, and so back pay calculations need not be precise; exactness is not expected") (internal quotations omitted); *Watson*, 2002 WL 31006129, at *8 ("Back pay relief is an equitable remedy, and the common law requirement of 'certainty' has never been applied to it."). When making back pay calculations, courts should err on the side of awarding greater damages rather than less damages to the plaintiffs because the "purpose of Title VII is to dissuade discrimination [and] awarding any windfall to the defendant may encourage, rather than discourage, discrimination in the future." *Artis v. U.S. Indus. & Int'l Ass'n of Machinists &*

Aerospace Workers, 822 F. Supp. 510, 511 (N.D. Ill. 1993); *see also Watson*, 2002 WL, at *9 ("ambiguities in what an employee would have earned but for the discrimination are to be resolved against the employer.").

251. Stacy Ernst was denied employment as a Paramedic for the Chicago Fire Department ("CFD") from 2005 through the present. In total, she would have earned $1,023,816.96 in wages and benefits at the CFD from 2005 through October 2014. Plaintiffs' Summary of Damages, attached as Exhibit A to Declaration of Scott G. Grimes Re Calculation of Plaintiff Damages ("Grimes Decl. Oct. 31, 2014, ECF No. 506-1"). Instead, Ms. Ernst was employed at several other organizations from 2005 through October 2014. Ernst Decl., Oct. 31, 2014, ECF No. 506-2 at ¶¶ 4-13. At those other organizations, Ms. Ernst earned $758,822.75 in wages and benefits. Grimes Decl. For the period of 2005 through October 2014, the difference between the wages and benefits that Ms. Ernst could have earned working at the CFD and the amount that she did earn is $276,781.16. *Id*. The prejudgment interest from the difference in these earnings is $40,532.40. *Id*. Thus, the total wages and benefits that Ms. Ernst lost due to discrimination is $317,313.55. *Id*. In addition, each month going forward, Ms. Ernst is losing $2,085.65 due to discrimination. *Id*.

252. Michelle Lahalih was denied employment as a Paramedic for the CFD from 2005 through October 2014. In total, she would have earned $1,023,816.96 in wages and benefits at the CFD from 2005 through October 2014. *Id*. Instead, Ms. Lahalih was employed at several other organizations from 2005 through October 2014. Lahalih Decl., Oct. 31, 2014, ECF No. 506-2 at ¶¶ 3-12. For the period of 2005 through October, Ms. Lahalih earned $667,167.62 in wages and benefits. Grimes Decl. The difference between the wages and benefits that Ms. Lahalih could have earned working at the CFD and the amount that she did earn is

$356,649.34. *Id*. The prejudgment interest from the difference in these earnings is $72,839.23. *Id*. Thus, the total wages and benefits that Ms. Lahalih lost due to discrimination is $429,488.57. In addition, each month going forward, Ms. Lahalih is losing $2,804.48 due to discrimination. *Id*.

253. Dawn Hoard was denied employment as a Paramedic for the CFD from 2005 through October 2014. In total, she would have earned $1,023,816.96 in wages and benefits at the CFD from 2005 through October 2014. *Id*. Instead, Ms. Hoard was employed in several other organizations from 2005 through October 2014. Hoard Decl., Oct. 31, 2014, ECF No. 506-3 at ¶¶ 3-12. At those organizations, Ms. Hoard earned $486,490.39 in wages and benefits. Grimes Decl. For the period of 2005 through October 2014, the difference between the amount Ms. Hoard could have earned in wages and benefits working at the CFD and the amount that she did earn in wages and benefits from the other jobs is $537,326.57. *Id*. The prejudgment interest from the difference in these earnings is $84,109.83. *Id*. Thus, the total wages and benefits that Ms. Hoard lost due to discrimination is $621,436.39. In addition, each month going forward, Ms. Hoard is losing $6,017.46 due to discrimination. *Id*.

254. Irene Res-Pullano was denied employment as a Paramedic for the CFD from 2005 through October 2014. In total, she would have earned $1,023,816.96 in wages and benefits at the CFD from 2005 through 2014. *Id*. Due to unlawful discrimination, Ms. Res-Pullano did not work at the CFD from 2005 through October 2014. Instead, Ms. Res-Pullano was employed at several other organizations from 2005 through October 2014. Res-Pullano Decl., Oct. 31, 2014, ECF No. 506-6 at ¶¶ 4-14. At those organizations, Ms. Res-Pullano earned $542,852.07 in wages and benefits. Grimes Decl. For the period of 2005 through October 2014, the difference between the wages and benefits that Ms. Res-Pullano could have earned at the CFD and the amount that she did earn in is $480,964.89. *Id*. The prejudgment interest from the difference in these earnings

is $74,688.45. *Id*. Thus, the total wages and benefits that Ms. Res-Pullano lost due to discrimination is $555,653.33. *Id*. In addition, each month going forward, Ms. Res-Pullano is losing $5,059.90 due to discrimination. *Id*.

255. Katherine Kean was denied employment as a Paramedic for CFD from 2005 through October 2014. In total, she would have earned $1,023,816.96 in wages and benefits at the CFD from 2005 through October 2014. *Id*. Instead, Ms. Kean was employed at several other organizations from 2005 through October 2014. Kean Decl., Oct. 31, 2014, ECF No. 506-4 at ¶¶ 3-12. At those organizations, Ms. Kean earned $762,095.32 in wages and benefits. Grimes Decl. For the period of 2005 through October 2014, the difference between the wages and benefits that Ms. Kean could have earned working at the CFD and the amount that she did earn is $278,360.61. *Id*. The prejudgment interest from the difference in these earnings is $49,042.60. *Id*. Thus, the total wages and benefits that Ms. Kean lost due to discrimination is $327,403.21.

256. Accordingly, based on the above findings of fact and conclusions of law, each plaintiff is entitled to the following amounts of back pay relief:

Stacy Ernst – $317,313.55
Dawn Hoard – $621,436.39
Katherine Kean – $327,403.21
Michelle Lahalih – $429,488.57
Irene Res Pullano – $555,653.33

Plaintiffs Proposed Findings of Fact and Conclusion of Law as to Lost Wage Evidence, ECF No. 506 at ¶ 13.

257. Plaintiffs are further entitled to ongoing monthly back pay for each month after October 2014 in which they did not receive either instatement or back pay. As of October 2014, these were the ongoing monthly damages for each Plaintiff:

Stacy Ernst – $2,085.65
Dawn Hoard – $6,017.46

Michelle Lahalih – $2,804.48
Irene Res Pullano – $5,059.90

Plaintiffs Proposed Findings of Fact and Conclusion of Law as to Lost Wage Evidence, ECF No. 506.

258. These ongoing monthly damages likely changed between October 2014 and the entry of these proposed findings of fact and conclusions of law. For example, as of January 2015, each Plaintiff would be entitled to a salary enhancement based on ten years in the CFD. PX 39A at ERN019141. Thus, each Party is directed to submit supplemental proposed findings regarding ongoing monthly damages and updating the calculation of prejudgment interest to the date of entry of the finding and conclusions.

259. In addition to back pay, each Plaintiff is entitled to instatement. Instatement is the preferred remedy under Title VII. *E.g.*, *Deloughery v. City of Chicago*, Case No. 02-cv-2722, 2004 WL 1125897, at *9 (N.D. Ill. May 20, 2004) *aff'd*, 422 F.3d 611 (7th Cir. 2005) (citing *Bruso v. United Airlines, Inc.*, 239 F.3d 848, 861 (7th Cir.2001)) (additional citation omitted). A defendant can overcome instatement's preferred status by demonstrating that instatement would result in a "working relationship fraught with hostility and friction." *Bruso*, 239 F.3d at 861 (citation omitted). Here, each Plaintiff met the City's minimum qualifications other than the PPT for the CFD paramedic position. Indeed, each Plaintiff met the pre-2000 requirements, which were a substantially equally valid selection procedure. Further, each Plaintiff's paramedic experience has shown that each Plaintiff before and after the test performed as a paramedic all of the twenty-four job tasks HPS designated as "essential." Tr. 570:1-572:7, Nov. 6, 2014. At the same time, the City failed to establish that instatement would lead to a relationship fraught with hostility and friction. Therefore, each Plaintiff is entitled to instatement to the CFD as a

paramedic with full seniority and benefits as they would have had they been hired in January 2005.

260. Lastly, Plaintiffs are entitled to attorney's fees and costs. 42 U.S.C. § 2000e-5(k). Plaintiffs are directed to submit their petitions for fees and costs within the time and in the manner provided in the Federal Rules of Civil Procedure and the Local Rules of the Northern District of Illinois.

CONCLUSION

There is no doubt that the paramedic job is and has always been crucial to the health and safety of the residents of the City of Chicago. Paramedics, including CFD paramedics, perform vital medical services. The PPT at issue in this case, however, has had an undeniable adverse impact on women, excluding them from paramedic positions in the CFD, without any demonstrated improvement in, or relationship to, public health or safety.

Plaintiffs have established that the City's use of the PPT has a severe disparate impact on women. Plaintiffs have proved systemic disparities of statistical and practical significance, made manifest by the fact that the PPT erects a barrier to employment for 40% of the women who apply for paramedic positions with the CFD but only 2% of the men. They have amply established a prima facie case of disparate impact.

In its defense, the City has failed to show that this disparate impact was job related or the result of business necessity. The City has failed to demonstrate a sufficient relationship between the tasks of a paramedic and the abilities it tested, or intended to test, with the PPT. The City has failed to take measures to ensure the reliability of the PPT—indeed failed to introduce any evidence of the reliability of the PPT. It has failed to demonstrate that the abilities tested by the PPT are job related. It has failed to show that the 935 cutoff score corresponds to any particular level of ability on any paramedic job task. It has failed to show that the abilities measured by the

PPT are required before entry into the Academy. Compounding these problems, the City has imposed an arbitrary cutoff score, unrelated to any ability level needed for the job. The City has failed to show that performance on the PPT is demonstrably related to, predictive of, or significantly correlated to performance on the job.

Even if the City has demonstrated job relatedness and business necessity, there were several less discriminatory alternatives. Both parties' experts agreed that less discriminatory alternatives existed. Moreover, the City's hiring practices before the PPT were also a less discriminatory alternative. For all the above reasons, judgment should be entered for the Plaintiffs.

Dated: January 29, 2015

Respectfully submitted,

s/ David Borgen

SUSAN P. MALONE
smalonelaw@sbcglobal.net
542 S. Dearborn Street, Suite 610
Chicago, IL 60605
(312) 726-2638

DAVID BORGEN
dborgen@gbdhlegal.com
BYRON GOLDSTEIN
brgoldstein@gbdhlegal.com
Goldstein, Borgen, Dardarian & Ho
300 Lakeside Drive, Suite 1000
Oakland, CA 94612
(510) 763-9800; (510) 835-1417 (Fax)

MARNI WILLENSON
marni@willensonlaw.com
542 S. Dearborn Street, Suite 610
Chicago, IL 60605
(312) 546-4910; (312) 261-9977 (Fax)

JOSHUA KARSH
jkarsh@hsplegal.com
Hughes, Socol, Piers, Resnick, & Dym
70 West Madison Street, Suite 4000
Chicago, IL 60602
(312) 604-2630; (312) 604-2631 (Fax)

Attorneys for Plaintiffs